Modèles d'evolution de proteines en environnement variable

Mathieu Hemery

THÈSE DE DOCTORAT
DE L'UNIVERSITÉ PIERRE ET MARIE CURIE

Spécialité : Physique

École doctorale : « Physique en Île-de-France »

réalisée

à Gulliver &
au Laboratoire Interdisciplinaire de Physique

présentée par

Mathieu HEMERY

pour obtenir le grade de :

DOCTEUR DE L'UNIVERSITÉ PIERRE ET MARIE CURIE

Sujet de la thèse :

Modèles d'évolution de protéines en environnement variable

soutenue le 21 octobre 2015

devant le jury composé de :

M.	Paul FRANÇOIS	Rapporteur
M.	Yves-Henri SANE-JOUAND	Rapporteur
M.	Luca PELITI	Examinateur
M.	Martin WEIGT	Examinateur
M.	Florent KRZAKALA	Directeur de thèse, membre invité
M.	Olivier RIVOIRE	Codirecteur de thèse

À AMBRE, L'EXEMPLE D'ÉVOLUTION ET D'APPRENTISSAGE LE PLUS SOURIANT DE CETTE THÈSE.

Remerciements

À l'heure d'écrire ces remerciements, mes premières pensées vont à ma femme, mon plus adorable soutien durant ses trois ans de thèse. À mes parents bien sûr pour m'avoir très tôt guider dans cette démarche toujours assoiffée de compréhension et infiniment curieuse du monde qui caractérise la recherche : « Pourquoi ci, pourquoi ça ? ». À mes grands parents, à mes frères et finalement au reste de ma famille ainsi qu'aux camarades de classe avec qui, sur les bancs et dans les cours, nous cherchions déjà à démonter et remonter le monde pour mieux le comprendre.

Je voudrais aussi accorder quelques lignes à l'ensemble de mes professeurs de physique, de math, de chimie et de toutes les autres matières depuis la maternelle jusqu'à l'université. Ils sont bien trop nombreux pour que je puisse accorder à chacun la place qui lui revient mais, si cette thèse contient un certain mérite, ils en sont les premières sources. Je l'ai remercie pour m'avoir offert une formation variée dont rien ne saurait être jugé superflu. C'est indubitablement à eux tous que je dois mon goût pour l'ensemble des sciences tant dans la recherche que dans l'enseignement ; qu'ils en soient tous félicités. J'espère pouvoir un jour passer à mes élèves cette belle vision du monde que j'ai reçu de vous.

Évidemment, je ne peux féliciter l'enseignement sans penser en premier lieu à mes directeurs de thèse : Olivier Rivoire & Florent Krzakala. Chacun a su, à sa manière m'encadrer tant sur le plan scientifique que personnelle. Ils m'ont aider à formaliser mes pensées et à découvrir de nouveau champs. Tout en me permettant de conserver cet équilibre entre une focalisation sur quelques questions précises et un certain vagabondage scientifique qui permet de garder cette large ouverture d'esprit et cette culture scientifique qui forment le cœur même de la recherche que j'aime.

Un grand merci enfin au groupe et au laboratoire qui m'ont accueilli à Grenoble, Sebastien, Bahram, Erik et Marc en particulier, mais plus généralement à l'ensemble des doctorants et des chercheurs du laboratoire. Ils ont su créer un cadre de recherche au sein duquel l'immense variété des thématiques traitées, de la spectroscopie à la motricité cellulaire m'a permis de jeter un bref coup d'œil sur des activités de recherche très différentes de miennes et d'attirer ainsi ma curiosité sur un vaste ensemble de sujets intrigants et passionnants.

Je voudrais finalement remercier les différents membres du jury

The most exciting phrase to hear in science, the one that heralds new discoveries, is not 'Eureka !' (I found it !) but rather, 'hmm.... that's funny...'
Isaac Asimov

pour avoir accepter d'assister à ma soutenance et l'ensemble des personnels administratifs des différentes institutions qui m'ont permis de mener cette thèse à son terme : les écoles doctorales de Polytechnique et de Paris ainsi que les laboratoires Gulliver et LIPhy.

Pour clore ces remerciement, je tenais à signaler que le Haut Commandement Karlan dément toute implication dans la rédaction de cette thèse.

à Saint-Martin-d'Hères, le 3 juillet 2015

Table des matières

Introduction

> La physique mathématique et la biologie moléculaire sont la poésie d'aujourd'hui. Ce sont elles qui traduisent et qui façonnent le monde et elles soulèvent chez les jeunes gens l'enthousiasme qui venait hier des poètes.
>
> Jean d'Ormesson *Un jour je m'en irai sans en avoir tout dit*, 2013.

Comprendre l'architecture du vivant est un problème à la fois passionnant d'un point de vue théorique et essentiel pour de nombreuses applications pratiques. En près de 200 ans, la biologie est passée du stade de parent pauvre des sciences à celui de champ de recherche bouillonnant et attirant les scientifiques de tous les domaines. Deux phénomènes sont au cœur de cette transformation. La théorie de l'évolution des espèces par sélection naturelle, tout d'abord, a révélé l'une des forces fondamentales qui gouvernent la dynamique des organismes biologiques. Dans un second temps, l'émergence de la biologie moléculaire a permis de manipuler directement les gènes des organismes étudiés, ouvrant ainsi une étude rigoureuse et quantitative du phénomène du vivant.

La biologie est intéressante à plus d'un titre. Premièrement, elle nous concerne tous et fait écho à des questions profondes qui passionnent l'humanité et les philosophes depuis bien longtemps même si nous sommes encore – très – loin de pouvoir y apporter des réponses. Deuxièmement, c'est un phénomène éminemment riche et complexe présentant une diversité et une précision admirable. Ce n'est pas un hasard si l'idée que tout ceci soit le simple produit du hasard est une idée difficile à appréhender. Que cette incroyable mécanique si finement réglée qu'est chaque organisme vivant : de la baleine à la bactérie, du séquoïa géant à l'algue monocellulaire puisse être expliqué par un phénomène aussi simple que la lutte pour la survie est assurément quelque chose de profondément perturbant.

Pour détourner la boutade d'un célèbre physicien : « Quiconque n'est pas choqué par la théorie de l'évolution ne la comprend pas ! ». Certes nous disposons aujourd'hui de tant de preuves qu'il n'est plus concevable de la contredire. Pourtant, sommes nous sûrs de l'avoir parfaitement comprise ? De nombreuses questions restent encore – et heureusement – à élucider avant de pouvoir dire que nous comprenons les mécanismes de l'évolution. Vous pensez que les variations entre un enfant et ses parents sont le résultats de mutations apparaissant au hasard dans la longue chaîne de son

ADN. Mais ces mutations sont-elles vraiment toutes aléatoires ? Vous pensez que la principale force guidant le destin des gènes est leur utilité pour la survie de leur organisme hôte. Mais la découverte de l'évolution neutre est venu chambouler cette idée simple. Vous pensez enfin que l'on peut représenter l'évolution comme un immense arbre dont chaque espèce vivant aujourd'hui serait une feuille… mais le réseau semble se croiser et se recouper si souvent qu'il vaudrait peut être mieux parler d'un buisson, voire… d'une forêt particulièrement dense et inquiétante.

Et c'est bien là l'un des points caractéristiques dans l'étude du vivant. Tout y semble toujours plus complexe, fourmillant d'exceptions et de cas particuliers si bien que seule une étude approfondie et minutieuse permet de déceler la loi générale tapie derrière les cas singuliers.

Cette richesse en fait évidemment un terrain extrêmement intéressant pour la recherche, bien loin de se limiter à la biologie car la complexité intrinsèque de la matière vivante demande l'aide de nombreuses compétences pour percer ses secrets. En première ligne, les physiciens et les chimistes travaillent depuis le départ avec les biologistes pour élucider les mécanismes du vivant. L'identification, la caractérisation et l'étude des différents éléments constitutifs du vivants : protéines, membranes, ADN, ont occupés et occupent encore les chercheurs de nombreuses disciplines.

Que ce soit à travers l'utilisation d'instruments complexes pour la cristallographie des protéines ou la compréhension de la thermodynamique hors équilibre d'une cellule vivante au cours de son existence, les compétences des physiciens forment un apport incroyable pour la biologie. Mais il faut bien voir là une richesse à double sens car les problèmes découverts au cœur de la biologie sont des problèmes neufs et toujours plus complexes qui forcent tous les scientifiques à pousser toujours plus loin leur expertise, à développer de nouveaux outils expérimentaux et théoriques et à comprendre toujours plus profondément leur propre discipline.

Cette thèse s'inscrit ainsi à la confluence de la physique statistique et de la biologie de l'évolution. Les systèmes vivants sont en effet souvent décrits selon trois approches différentes et complémentaires : biologique, physique et évolutionniste ; et nous pensons que ces deux disciplines sont particulièrement indiquées pour rassembler ces trois aspects.

Pour le biologiste, c'est la fonction du système qui est l'objet d'étude. Quel est le rôle des différentes composantes de ce dernier ? Comment les différentes pièces du puzzle s'imbriquent elles les unes dans les autres ? Quelle est sa réponse à différentes sollicitations extérieures ? Si nous prenons l'exemple d'une protéine, ces molécules élémentaires de la plupart des organismes, le biologiste cherchera à comprendre si elle opère de façon isolée ou si elle peut se lier à une cible donnée ? Le cas échéant quelles sont ses cibles,

qu'en fait-elle, active t-elle une réaction chimique et pourquoi ? etc.

De son côté, le physicien va plutôt chercher à comprendre comment le système remplit sa fonction, par exemple quelles sont les forces physiques fondamentales pour cette dernière. Pour reprendre l'exemple des protéines, on se demandera comment cette dernière se replie pour devenir fonctionnelle ? Quels atomes sont essentiels pour activer la réaction ? Ou encore comment une protéine donnée peut-elle identifier avec une très haute précision une cible particulière parmi des milliers d'autres ?

Enfin, l'approche évolutive va chercher à apporter un éclairage historique. Pourquoi telle protéine ressemble t-elle à telle autre ? Comment peut-on identifier les éléments les plus importants sans même connaître la physique de la protéine ? Elle cherche aussi à comprendre plus quantitativement pourquoi une fonction sera préférée à une autre et, plus important encore, elle permet d'expliquer l'architecture du vivant c'est-à-dire les raisons qui dictent la forme des différentes pièces du puzzle.

Actuellement, lorsque l'on parle d'expliquer une fonction, on entend surtout une approche physique, c'est-à-dire comment les règles de la physique permettent de rendre compte des observations biologiques. Notre approche repose sur le constat qu'il existe, pour une fonction donnée, un grand nombre de solutions différentes, mais que les solutions naturelles non seulement n'en représentent qu'un petit sous ensemble, mais surtout un sous ensemble très particulier. La question que nous nous posons est donc de comprendre les propriétés de cet ensemble au regard de l'évolution. Autrement dit, comment et pourquoi l'histoire évolutive d'un système le contraint-il ?

Pour rendre plus claire cette notion de contrainte dans l'espace des solutions, revenons un instant à une description plus physique. Un système biologique comprend généralement un grand nombre d'éléments en interaction les uns avec les autres. De ces interactions émergent un certains nombre de propriété. Une propriété donnée peut dépendre de l'ensemble des interactions ou seulement d'un petit nombre d'entre elles. Par exemple, la fonction d'une protéine pourrait être sensible à la modification de n'importe quel acide aminé ou au contraire reposer seulement sur un petit nombre d'entre eux. De même la résistance d'une bactérie à un antibiotique pourrait dépendre de l'organisation de l'ensemble de ses gènes ou bien de seulement quelques uns exprimés au bon moment. Il s'avère que très souvent, c'est le second choix qui est effectué : la propriété dépend essentiellement d'une partie restreinte et structurée de l'ensemble du système, phénomène que nous appellerons parcimonie.

C'est cette architecture, cette géométrie particulière que nous avons à l'esprit lorsque l'on dit que les solutions naturelles (séquence d'acide aminé, motif d'un réseau de gènes, etc.) ne sont

donc pas uniformément réparties dans l'ensemble des solutions possibles. Ceci pose inévitablement de nombreuses questions. Quels sont les biais et peut-on les quantifier ? Quelles sont les forces de l'évolution responsable de l'apparition de telle ou telle solution ? L'évolution peut-elle favoriser des propriétés comme la capacité à modifier la fonction, ou à préserver la fonction, voire les deux en même temps ? Qu'est ce que l'étude d'une fonction aujourd'hui peut nous apprendre sur le passé ?

Pour répondre à toutes ces questions, nous avons développé un modèle minimal permettant, dans un cadre épuré, de tester nos hypothèses et de faire ressortir les paramètres clés pour l'apparition de propriétés particulières. Nous nous concentrerons dans le reste de ce tapuscrit sur *l'étude des liens existants entre la géométrie d'une protéine et son histoire évolutive*.

Avant de présenter nos travaux et afin de permettre au lecteur de comprendre la source et la portée de ces derniers, nous commencerons par une présentation détaillée des protéines, à la croisée de la biologie et de la physique. Nous détaillerons différents modèles inspirés de la physique ayant permis de mieux comprendre le fonctionnement et la nature de ces molécules complexes. Nous présenterons ensuite la théorie synthétique de l'évolution tant dans son aspect naturel que d'un point de vue purement algorithmique. Nous expliquerons ensuite comment cette dernière permet une lecture plus riche et plus claire du vivant en général et des protéines en particulier. Par exemple à travers l'utilisation d'algorithmes inspirés de l'évolution sur des modèles de protéines. Nous montrerons enfin comment des modèles de physique statistique appliqués aux séquences des protéines permettent de révéler d'une manière originale l'information présente dans ces dernières. Ceci permet d'accéder ainsi à un éclairage intéressant sur l'architecture des protéines et sur le fonctionnement de l'évolution qui vient compléter les points de vue structurels et fonctionnels.

Ces bases expliquées, nous présenterons rapidement nos résultats et leurs implications. Nous développerons ensuite un modèle jouet permettant de mettre en avant les éléments centraux de nos travaux, avant de détailler de manière exhaustive le modèle plus complet que nous avons choisi d'utiliser. Nous évoquerons aussi ses différentes variantes afin de souligner la généralité de nos résultats. Nous montrerons ensuite comment ce modèle permet de rendre compte et d'expliquer les différentes propriétés observées dans la géométrie des protéines en permettant de reproduire dans un cas simple les différents points de vue utilisés pour étudier ces dernières.

Nous terminerons en présentant en quelques lignes les différentes manières dont l'on pourrait poursuivre les thématiques développées dans ce tapuscrit ainsi que quelques appendices destinés à ceux désirant comprendre cette thèse plus en détail.

Première partie

Position & contexte

Les Protéines : Objet biologique, Objet physique

> Le corps humain contient 60 000 milliards de cellules et dix fois autant de bactéries. Environ. Et on se laisse encore impressionner par le nombre d'étoiles dans le ciel.
>
> Francis Dannemark

Objet central de la biologie moléculaire, les protéines sont les principales molécules organiques présentes dans les cellules. Leurs fonctions couvrent pratiquement l'ensemble des opérations réalisées au sein d'un organisme, depuis le mouvement de l'individu jusqu'à la synthèse… des protéines.

D'un point de vue chimique, une protéine est un hétéro-polymère linéaire constitué d'une simple chaîne dont les blocs – ou monomères – sont appelés acides aminés. Si la chaîne est courte, on préférera parler de peptide, réservant le terme de protéine pour les chaînes de plus de 50 acides aminés.

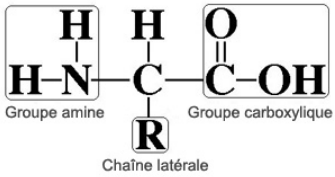

Figure 1: Structure générale d'un acide aminé.

Chaque acide aminé est constitué de trois éléments : une amine, un carbone central, ou carbone α, qui porte la chaîne latérale aussi appelé résidu et un groupement carboxyle. La polymérisation est constituée de la formation d'un amide entre l'amine d'un acide aminé et le groupe carboxyle du suivant. Comme seul le résidu varie d'un élément à l'autre, c'est ce dernier qui sert à classifier la vingtaine d'acides aminés présents dans l'ensemble du vivant. Ces derniers sont traditionnellement indiqués par des lettres latines (généralement en capitale) et répartis en différentes catégories selon leurs propriétés physico-chimiques comme illustré dans la figure (2).

La nature des acides aminés le long de la chaîne constitue la *structure primaire* de la protéine. Cette dernière est déterminée lors de la *traduction* de la séquence de l'acide désoxyribonucléique (ADN) du gène associé à la protéine par l'intermédiaire du code génétique qui associe un acide aminé à chaque triplet de nucléotides [1]. Le lien entre la séquence de l'ADN et celle de la protéine peut cependant s'avèrer plus complexe car certaines protéines peuvent être modifiées après avoir été traduites, une étape nommée la *modification post-traductionnelle*.

1. À de rares exceptions près, chaque triplet code pour un acide aminé, plusieurs triplets peuvent cependant coder le même acide.

Le paradigme dominant aujourd'hui la biologie veut alors que la séquence détermine le repliement de la protéine qui va à son tour gouverner la fonction de cette dernière. La caractérisation de la structure primaire fut donc une réalisation majeure de la biologie au cours du XXème siècle. Elle débuta avec la découverte de l'ADN

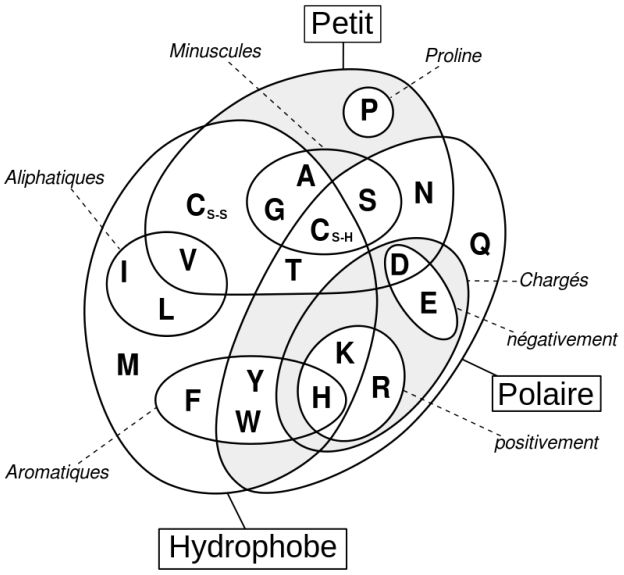

FIGURE 2: Diagramme de Venn indiquant les propriétés physico-chimiques des différents acides-aminés.

par WATSON et CRICK [2], se poursuivit avec le décryptage du code génétique et continue aujourd'hui avec le séquençage du génome de nombreux organismes dont l'homme au début des années 2000 [3]. La connaissance de la structure primaire c'est cependant révélée loin de fournir l'ensemble des réponses désirées et un travail considérable reste encore à fournir afin d'identifier et de déterminer les informations contenues dans les protéines.

En effet, une fois la séquence connue, il est nécessaire de déterminer la forme repliée de la protéine puisque c'est dans cette dernière que la majorité des protéines accomplissent leur fonction [4]. Lors de ce repliement on identifie d'abord deux types de structure locales : les hélices α et les feuillets β qui forment la *structure secondaire* de la protéine. Ces éléments se replient ensuite dans l'espace pour former la *structure tertiaire*. Lorsque plusieurs protéines se lient les unes aux autres pour former des structures plus larges, les multimères, on parle alors de *structure quaternaire*.

Il est courant, notamment dans le cas des grandes protéines ou lors d'un choc thermique, que le repliement ne puisse s'effectuer de manière autonome. D'autres protéines, appelées protéines chaperons [5], viennent alors assister ces dernières [6]. Notons cependant qu'il n'existe pas toujours une conformation unique pour chaque protéine, certaines protéines remplissant même leur fonction directement depuis leur état dépliée, on parle alors de protéines intrinsèquement désordonnées [7].

Sous-éléments de la structure tertiaire, les *domaines* de protéines sont des sous ensembles stables de la protéine souvent capable de se replier de manière autonome et parfois porteurs de fonctions spécifiques. Les protéines sont souvent constituées, de façon modulaire, de plusieurs de ces sous-unités. Certains domaines peuvent être retrouvés au sein de nombreuses protéines au sein du même organisme ou dans des organismes différents, on parlera de do-

2. F Crick and J Watson. Molecular structure of nucleic acids. *Nature*, 1953

3. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature*, 2004

4. C'est le cas pour 50 à 70% des protéines chez les eukaryotes et de 75 à 95% chez les archées et les bactéries.

A K Dunker, P Romero, Z Obradovic, and E C Garner. Intrinsic protein disorder in complete genomes. *Genome Informatics*, 2000

5. R J Ellis and S M Hemmingsen. Molecular chaperones: proteins essential for the biogenesis of some macromolecular structures. *Trends in biochemical sciences*, 1989

6. Certaines protéines sont aussi chargées de dégrader les protéines s'étant mal repliées afin d'éviter toute réaction parasite.

7. H Jane Dyson and Peter E Wright. Intrinsically unstructured proteins and their functions. *Nature Reviews Molecular Cell Biology*, 6(3):197–208, March 2005

maines « homologues » puisqu'ils possèdent une origine commune.

Les fonctions réalisées par les protéines (ou les domaines de protéine) sont très variées, les principales étant :

Liaison : la capacité à se fixer de manière efficace et spécifique à des cibles particulières (les ligands) est essentielle pour de nombreuses protéines, souvent couplée à d'autres fonctions telle que l'activité enzymatique ou la régulation. Elle peut aussi servir au marquage d'élément dans la cellule ou au transport de molécules au sein des organismes *(Exemple : anticorps, hémoglobine, activateurs et répresseurs de gènes)*

Enzyme : désigne les catalyseurs organiques. Ces derniers sont capables d'accélérer des réactions chimiques jusqu'à la limite de diffusion [8]. Les protéines enzymatiques sont au cœur de la plupart des processus du vivant, puisqu'elles rendent la vie possible dans des échelles de temps raisonnables *(Exemple : β-lactamase, digestion)*

8. On parle alors d'enzymes parfaites.

Structure : que ce soit en permettant à la cellule d'échanger avec l'extérieur afin de réguler sa composition interne, en s'accrochant au substrat extérieur ou en formant un endo-squelette rigide, les protéines participent à la forme et aux propriétés mécaniques de la cellule *(Exemple : protéine membranaire, actine)*

Mobilité : du mouvement collectif des fibres de myosines dans les muscles des vertébrés aux flagelles des bactéries, les protéines sont les principales responsables des mouvements macroscopiques des organismes comme des cellules individuelles *(Exemple : actine, myosine, moteur moléculaire)*

Régulation : que ce soit en contrôlant les flux au niveau de la membrane ou en maîtrisant la production et la dégradation des différentes protéines, les protéines jouent un rôle central dans la composition chimique de la cellule et dans les mécanismes de réponse de cette dernière *(Exemple : protéine membranaire, recruteur, répresseur)*

Expression génétique : les protéines sont responsables de la réplication de l'ADN ainsi que de sa traduction en protéines. Ces fonctions clés du vivant sont parmi les éléments les mieux conservés du génôme entre tous les êtres vivants *(Exemple : Ribosome, ADN Polymerase)*

Enfin, il faut garder à l'esprit que le nombre, mais aussi les proportions des différentes protéines présentes dans une cellule dépendent entre autres : de la nature de cette dernière (neurones, cellules de la peau, des muscles, du foie, etc.) des conditions exterieures (présence de glucose, de lumière, etc.) et intérieures (effet de mémoire). Comprendre quelles sont les conditions pour lesquelles chaque protéine est produite et dégradée par la cellule constituent l'un des objectifs fondamentaux de la biologie cellulaire.

La production de protéine est en effet la principale dépense d'énergie d'un organisme [9], si bien que leur production est finement régulée par un ensemble de protéines, de fragments d'ADN,

9. J B Russell and G M Cook. Energetics of bacterial growth: balance of anabolic and catabolic reactions. *Microbiological reviews*, 1995

d'ARN et de gènes qui forment ce que l'on nomme un *réseau de gènes*. L'exemple le plus connu est celui de l'opéron lactose qui contrôle entre autre la production d'une enzyme permettant de dégrader le sucre lactose. Cette dernière n'est toutefois produite par la cellule qu'en présence de lactose... et en l'absence de glucose dont l'assimilation par la cellule est plus efficace.

Dans ce domaine, l'utilisation de la technologie RNA-seq [10] permet d'obtenir un suivi temporel de la transcription d'un gène donné. Ceci permet de comprendre quand un gène est activé et surtout, quels gènes sont transcrits de concert. L'ensemble des ARN traduits simultanément est désigné par le terme de *transcriptome* et dépend de l'organisme, de la cellule étudiée et des conditions externes.

10. Z Wang, M Gerstein, and M Snyder. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet*, 2009

Pour une bactérie ordinaire comme *Escherichia coli* [11], les protéines représentent environ la moitié de la masse sèche. C'est donc indubitablement l'un des objets d'étude les plus important de la biologie au vu de son omniprésence et de sa grande polyvalence au sein de la cellule. Pour autant, le terme de protéine recouvre une telle diversité d'objets qu'il est sans doute illusoire de vouloir construire une unique théorie pour décrire des objets aussi différents que peuvent l'être les anticorps et les flagelles.

11. *Escherichia coli* est une bactérie présente dans l'intestin des mammifères. Elle est particulièrement étudiée par la communauté scientifique. Certaines souches peuvent être pathogènes.

Des modèles simples sont cependant possibles pour décrire le comportement des protéines dans leur ensemble (pour le repliement par exemple) ou dans des cas plus précis (modèle de liaison d'anticorps pour l'immunologie, etc.). Chacun permet d'éclairer un aspect essentiel des protéines et de mettre en valeur ce qui relève du comportement ordinaire d'une molécule et ce qui mérite une attention renouvelée pour en comprendre les mécanismes particuliers. De nombreux physiciens se sont donc interessés aux différents problèmes que posent les protéines afin d'apporter, par leurs méthodes et leurs techniques, un éclairage intéressant sur ces questions complexes.

On peut citer deux grandes manières d'approcher le problème : déterminer les propriétés générales d'un système vivant ou bien proposer un modèle simple permettant de comprendre un objet ou une propriété en particulier.

La première approche tient sa source dans le principe même de la physique statistique qui consiste à s'occuper non pas d'un élément particulier pour lequel le hasard tient une place prépondérante, mais au contraire de s'intéresser au comportement moyen d'un très grand nombre d'éléments, afin que les éléments aléatoires se compensent, permettant de dégager les propriétés et les quantitées essentielles du système.

Prenons l'exemple physique d'un verre d'eau : décrire une molécule est chose aisée, en décrire deux est difficile et cent pratiquement impossible. Heureusement, le verre en contenant de l'ordre de 10^{25}, tout se simplifie et des quantités nouvelles comme l'énergie,

l'entropie, la pression, le volume suffisent pour décrire avec une grande précision le comportement macroscopique du liquide.

Quelles situations similaires peut-on trouver en biologie ? Décrire le comportement d'une protéine particulière est un travail harassant, en décrire dix est titanesque mais peut-on formuler des lois décrivant de très nombreuses protéines en interaction ? Peut-on estimer le comportement des milliards de bactéries présentes dans une colonie ? Peut-on prédire l'évolution de millions d'individus ? L'espoir est bien de trouver des quantités et des propriétés nouvelles permettant de quantifier la biologie au-delà d'une certaine échelle.

La seconde méthode permet de mettre en valeur une propriété particulière et de fournir une représentation pratique au biologiste pour concevoir de nouvelles expériences. Elle présente donc l'intérêt d'apporter des réponses plus directes à des problèmes concrets mais risque aussi d'être confrontée à un grand nombre de cas particuliers au vu de la nature pour le moins hétérogène de la matière biologique.

Nous présenterons dans la suite de ce chapitre quelques exemples où l'approche physique a permis de comprendre certains aspects d'un problème biologique. Que ce soit en en déterminant les paramètres clés ou en les simulant qualitativement, à l'aide de modèles simplifiés. Nous nous concentrerons d'abord sur les techniques de détermination expérimentale de la structure tertiaire avant de détailler un modèle permettant d'estimer le diagramme des phases du repliement. Nous discuterons enfin des problèmes au delà de la structure tertiaire.

Determination expérimentale de la structure tertiaire

Connaître en détail la structure d'une protéine constitue souvent un véritable défi technologique et un enjeu de taille pour les biologistes. Les deux principales méthodes actuellement utilisées sont la *cristallographie* et la *spectroscopie RMN*.

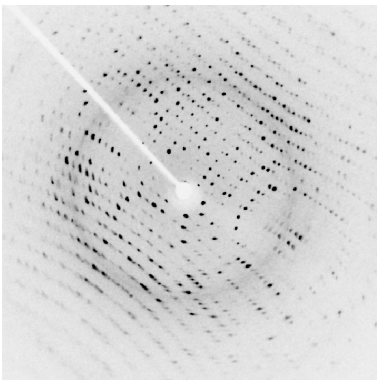

FIGURE 3: Cliché de diffraction par rayon X d'un cristal de lysozyme, une enzyme présente entre autre dans le blanc d'œuf. *Image publiée sur Wikimedia par Del45.*

Ces deux méthodes complémentaires présentent cependant de nombreuses questions quand à la validité de leurs résultats. La cristallographie, par exemple, demande de produire un cristal de protéines, ce qui n'est possible qu'à de très fortes concentrations et à des température inférieures à celles présentes dans le vivant. La spectroscopie travaille quand à elle sur des échantillons très purifiés et relativement dilués. Dans ces deux cas, les conditions sont très éloignées de celles présentes dans la cellule vivante et la conformation que la protéine adopte alors pourrait être très éloignée de sa forme fonctionnelle, même si de nombreux indices laissent à penser que les informations ainsi récoltées sont exactes dans une large mesure. De plus, certaines protéines peuvent avoir plusieurs conformations correspondant par exemple à des formes actives ou inactives ; ou encore rester dépliées au sein de la cellule, ce qui élimine la cristallographie et rend la spectroscopie difficile.

La détermination de la structure tri-dimensionnelle n'est ce-

pendant pas la fin de l'histoire. En effet, les protéines remplissent différent rôles au sein de l'organisme et pouvoir expliquer et quantifier le rôle de chaque protéine demande un travail plus important encore. Il faut en effet comprendre la façon dont la protéine remplit cette fonction. Par exemple, dans le cas d'une enzyme, quel est son site de liaison, quels acides aminés sont nécéssaires à l'activité de réaction, y a t-il un changement de conformation, etc.

Cristallographie : Diffraction par rayon X

La diffraction par rayon X (DRX) est l'une des principales méthodes de cristallographie utilisées pour obtenir des informations sur la structure tridimensionnelle des protéines. Dans la pratique, on commence par préparer un cristal de protéine en refroidissant un échantillon purifié d'une solution à forte concentration de protéines. Puis on éclaire l'échantillon par une très brève impulsion d'un rayonnement de petite longueur d'onde qui va, par diffraction donner des informations sur la structure de la protéine via le facteur de forme. L'impulsion doit être brève car l'échantillon chauffe rapidement ce qui provoque la fonte du cristal et brouille rapidement la mesure. Ceci signifie aussi que la mesure est destructive et que chaque mesure demande de produire un nouveau cristal.

Cette méthode permet d'accéder en premier lieu à la symétrie de la protéine, mais nécessite ensuite un important travail d'analyse pour remonter à la structure et enfin au modes de vibration de cette dernière. En particulier, la détermination du facteur B (ou facteur de Debye-Waller) lié au bruit thermique permet dans le cas des protéines de déterminer les fluctuations de chaque résidu autour de sa position d'équilibre et donc d'accéder à la stabilité de la protéine ainsi qu'à ses principaux modes et mouvements. Ces informations peuvent alors être utilisées pour obtenir une idée plus précise de la structure, mais fournissent aussi des indices sur la façon dont la protéine peut réaliser sa fonction.

Spectroscopie par résonance magnétique nucléaire

La spectroscopie par résonance magnétique nucléaire (RMN) permet d'obtenir des informations sur l'environnement électronique des protons le long de la séquence d'acides aminés. On construit ainsi petit à petit la structure en reliant la séquence connue aux informations spectrales tirées de la spectroscopie. Cette méthode, contrairement à la cristallographie, étudie donc directement la structure locale de la protéine. Dans le cas de protéines de grandes tailles où de nombreux protons présentent un environnement similaire, on aura parfois recours à un marquage isotopique sur certaines positions pour pallier à cet écueil.

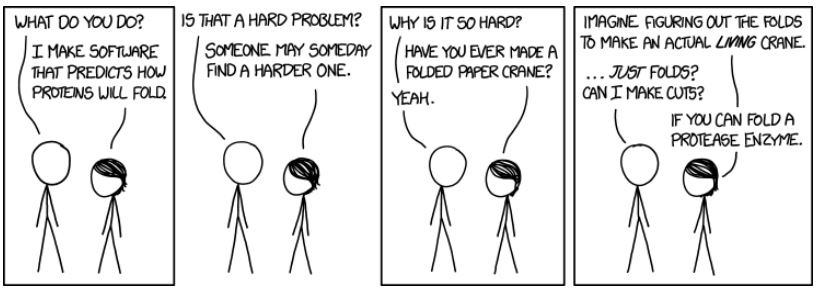

FIGURE 4: Proteins, XKCD.

Le problème du repliement

Le premier problème qui vient à l'esprit d'un physicien lorsque l'on lui parle de protéine est celui du repliement. En 1970, LEVINTHAL, un biologiste américain, fit remarquer qu'au vu du nombre de conformations possibles d'une chaîne d'acides aminés, il faudrait normalement des milliards d'années à une protéine pour trouver l'état d'énergie minimale, même en testant ses possibilités à un rythme effréné [12]. Cette remarque suggère que les protéines se replient suivant une séquence particulière, intuition issue de la mise en évidence expérimentale de l'auto-repliement de petites protéines dans des temps bien plus court que ceux prédits ci-dessus [13].

Cette observation est loin d'être anecdotique, puisque la séquence temporelle suivant laquelle ces molécules se replient est elle-même encodée dans la séquence d'acides aminés. En déchiffrant cette information, il devrait donc être possible de suivre ce chemin et de prédire ainsi la structure tridimensionnelle d'une protéine à partir de sa séquence sans avoir à parcourir exhaustivement l'espace des configurations – ce qui, on l'a vu, est impossible.

D'un point de vue de physicien, la première interrogation porte sur la nature des interactions médiant le repliement, c'est-à-dire quelles sont les interactions essentielles à ce dernier [14]. Le plus grand rôle est-il joué par les interactions de courte portée ou de longue portée et quels sont les effets du solvant par exemple. La deuxième interrogation porte sur la structure hiérarchique de cette dynamique. La protéine se forme t-elle tout d'un coup, ou progressivement, au fur et à mesure de sa synthèse, ou bien encore par groupes successifs – par exemple domaines par domaines, ou encore en suivant par exemple d'abord la structure secondaire puis la structure tertiaire.

Nous présenterons pour ces questions d'abord une approche plutôt force-brute : la dynamique moléculaire, puis un modèle jouet datant du début des années 1990 suivant une approche de physique statistique originale et intéressante.

12. Les deux angles d'un acide aminé produisent typiquement 3 configurations acceptables. Une chaîne de 100 AA possède donc environ 3^{100} conformations. À un rythme d'une toutes les femtosecondes, il ne nous reste plus qu'à attendre 10^{18} fois l'âge de l'univers.

13. C B Anfinsen and E Haber. Studies on the reduction and re-formation of protein disulfide bonds. *J Biol Chem*, 1961

14. C B Anfinsen. Principles that govern the folding of protein chains. *Science*, 1973

Dynamique Moléculaire

L'approche la plus naturelle pour étudier la dynamique de repliement des protéines fait appel à la Dynamique Moléculaire

(DM), c'est-à-dire la description déterministe et plus ou moins détaillée de la molécule et de ses différentes interactions afin d'en déduire les forces s'exerçant sur les différentes parties et d'en prédire le mouvement.

Pour commencer, il faut choisir un niveau de description de la protéine allant de la simulation complète : la simulation tous-atomes (où chaque atome est représenté avec ses liaisons et ses interactions), jusqu'au modèle fortement décimé utilisant un seul élément pour chaque acide aminé, voire moins. Cette méthode est donc confrontée à un choix cornélien entre la volonté d'obtenir des informations réalistes et extrêmement détaillées d'une part et les possibilités de calculs envisageables d'autre part.

Dans les deux cas, notons cependant que ce type de simulation demande de connaître de façon quantitative l'ensemble des interactions élémentaires de la protéine, c'est-à-dire les interactions des atomes ou des acides aminés entre eux par exemple, mais aussi l'influence du solvant et des éventuelles ions présents, etc. La détermination exacte de ces paramètres est déjà un véritable défi technique en soi [15].

Dans le cas où l'on chercherait une simulation détaillée, comme une protéine est un objet extrêmement complexe de plusieurs milliers d'atomes, cette approche demande donc une puissance de calcul phénoménale pour obtenir des temps simulés relativement courts. Des simulations tous-atomes atteignent aujourd'hui des temps physiques de l'ordre de la centaine de micro seconde dans le meilleur des cas [16] pour des temps de calcul se comptant en mois. Notons toutefois que ces échelles de temps correspondent déjà à celles nécéssaire pour l'étude de nombreuses fonctions biologiques.

Même si d'un point de vue technique il est toujours possible d'attendre l'apparition de calculateurs plus puissants ou de construire des processeurs dédiés spécifiquement à cette tâche, comme Anton construit par le groupe de Shaw [17], il est souvent préférable d'adopter des ambitions plus restreintes, par exemple en représentant les acides aminés de façon fortement décimée par un ou deux objets plutôt que d'en décrire le détail atomique. On peut ainsi simuler des temps physiques beaucoup plus longs mais l'interprétation des résultats demande évidemment un peu plus de précautions.

Dans tous les cas, et tant que la puissance de calcul restera une ressource limitante, les deux approches resteront complémentaires pour comprendre la dynamique des protéines.

15. Stefano Piana, Kresten Lindorff-Larsen, and David E Shaw. How Robust Are Protein Folding Simulations with Respect to Force Field Parameterization? *Biophysical Journal*, 100(9):L47–L49, May 2011

16. Stefano Piana, Alexander G Donchev, Paul Robustelli, and David E Shaw. Water Dispersion Interactions Strongly Influence Simulated Structural Properties of Disordered Protein States. *J. Phys. Chem. B*, 119(16):5113–5123, April 2015

17. David E Shaw *et al.* Anton, a special-purpose machine for molecular dynamics simulation. *Commun. ACM*, 51(7):91, July 2008

Verre de spin

Si la dynamique moléculaire permet de décrire le comportement d'une protéine en particulier, de nombreuses approches plus théoriques permettent de tirer des conclusions générales sur les problèmes étudiés.

Un exemple de modèle cherchant à déterminer les caractéris-

tiques saillantes du problème du repliement plutôt que les détails précis fut présenté par BRYNGELSON et WOLYNES [18]. Cette approche est d'autant plus intéressante pour nous qu'elle étudie la matière organique comme un nouvel état de la matière, à mi-chemin entre l'ordre et le désordre, ce qui n'est pas sans rappeler le problème de la transition vitreuse.

18. Joseph D Bryngelson and Peter G Wolynes. Spin glasses and the statistical mechanics of protein folding. *PNAS*, 84(21):7524–7528, 1987

Les auteurs y utilisent justement un modèle inspiré des verres de spin afin de déterminer l'allure du diagramme des phases d'une protéine se repliant. Il s'agit en particulier de déterminer les grandeurs physiques qui contrôlent ce processus et d'en déduire les conditions pour que la protéine se replie ou non et l'unicité de ce repliement.

L'idée générale est de restreindre la description de la protéine à un nombre limité de variables discrètes. Ici, on suppose une variable à ν états pour chaque acide aminé en choisissant ν de façon à reproduire les mesures expérimentales de différence d'entropie : $\nu \simeq 10$. Le hamiltonien comporte alors trois termes :

— un terme local déterminant l'énergie de chaque conformation,

— un terme de chaîne décrivant les interactions le long de la séquence,

— un terme à longue portée pour les interactions entre acides aminés éloignés.

Formellement on a donc :

$$E(\sigma) = -\sum_i \epsilon_i(\sigma_i) - \sum_i J_i(\sigma_i, \sigma_{i+1}) - \sum_{i,j} K_{ij}(\sigma_i, \sigma_j, r_i, r_j), \qquad (1)$$

où σ décrit la conformation de la protéine et ϵ_i, J_i et K_{ij} forment des constantes énergétiques *a priori* inconnues. Du choix de ces constantes dépendra le paysage énergétique ressenti par la protéine est donc son repliement : ces dernières sont donc cruciales. Il est tout simplement hors de question de chercher à en donner une estimation réaliste tant ces variables sont nombreuses et connectées entre elles. Nous pouvons cependant adopter un point de vue aléatoire inspiré du REM [19], en ne différenciant pour toutes ces interactions que deux cas : les interactions natives, c'est-à-dire celles dont toutes les variables sont dans l'état de la conformation finale supposée unique [20], de basse énergie $-\epsilon$, $-J$, $-K$ et les interactions non-natives ayant une énergie gaussienne aléatoire de moyenne $-\overline{\epsilon}$, $-\overline{J}$, $-\overline{K}$ et de variance $\Delta\epsilon$, ΔJ, ΔK.

19. Le modèle d'énergie aléatoire (*Random Energy Model* en anglais) est l'un des modèles de verre de spin les plus simples, chaque configuration ayant une énergie différente tirée aléatoirement. Il présente cependant déjà une transition de type vitreux ce qui en fait tout son intérêt.

Bernard Derrida. Random-energy model: An exactly solvable model of disordered systems. *Physical Review B*, 24(5):2613, 1981

L'énergie E d'une configuration quelconque devient alors une variable aléatoire gaussienne ne dépendant que du nombre N_0 de variables dans l'état natif que l'on supposera réparties aléatoirement le long de la séquence. La distribution de l'énergie des conformations présentant N_0 variables dans leur état natif est dans ce cas une gaussienne de moyenne :

$$\overline{E}(N_0) = -N_0\epsilon - \frac{N_0^2}{N}L - (N - N_0)\overline{\epsilon} - (N - \frac{N_0^2}{N})\overline{L} \qquad (2)$$

20. Le terme de natif provient des expériences sur le repliement des protéines. Il faut en effet isoler une protéine repliée (c'est donc là son état natif), puis la dénaturer afin d'observer son repliement vers la configuration stable dont on l'a tirée.

et d'écart-type :

$$\Delta E(N_0) = \sqrt{(N - N_0)\Delta\epsilon^2 - (N - \frac{N_0^2}{N})\Delta L^2}. \quad (3)$$

Ici, L est une variable regroupant les interactions à deux points provenant du deuxième et troisième terme de (1) est donnée par :

$$\begin{aligned} L &= J + zK, \\ \overline{L} &= \overline{J} + z\overline{K}, \\ \Delta L^2 &= \Delta J^2 + z\Delta K^2, \end{aligned} \quad (4)$$

où z quantifie le nombre de contacts moyen de chaque acide aminé.

L'énergie de l'état natif étant $\overline{E}(N) = -N(\epsilon + L)$, on peut désormais répondre à deux questions. Le système est il dans un état gelé – d'entropie nulle – ce qui signifie que la dynamique du système s'est arrêtée et qu'il est bloqué dans une unique configuration (fusse-t-elle la bonne) ? Existe t-il des conformations différentes d'énergie plus basse que la conformation native, autrement dit, le système s'est-il replié correctement ?

La figure (5) présente un exemple typique du diagramme de phase d'un tel système où l'on peut voir les quatre phases correspondant aux différentes réponses possibles. Dans les phases "gelées", l'entropie du système est nulle et il n'y a plus d'exploration des différentes configurations [21]. Dans la partie "non-repliée", la majeur partie des protéines sont dans un état différent de la conformation native tandis que dans les parties "repliée" cette conformation est majoritaire au sein de la population.

On voit que le système est donc dominé par deux paramètres [22] : la variance des conformations non-natives ΔL et l'écart énergétique entre les conformations natives et non-natives $L - \overline{L}$. Ce paramètre aussi appelé *gap* énergétique joue donc un rôle central dans la théorie du repliement. Essentiellement, on peut estimer qu'une protéine se replie dans sa conformation native efficacement dès lors que le *gap* est suffisamment important [23]. On voit de plus que même en l'absence de variance, le coût entropique pour replier la molécule implique un écart certain entre l'énergie des états natifs et non-natifs et que dans notre cas les transitions de nature entropique sont du second ordre.

Ce modèle apporte donc un éclairage sur les questions de conceptions de protéines en montrant quel sont les paramètres sur lesquels on peut jouer pour obtenir des configurations stables et quelles sont les principales forces qui entrent en jeu dans le repliement.

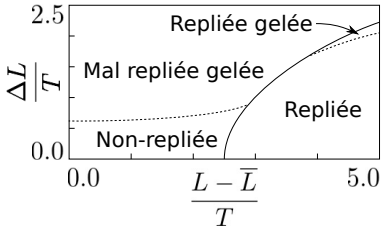

FIGURE 5: Exemple de diagramme de phase du modèle de verre de spin, les traits pleins correspondent à des transitions du premier ordre, les traits en pointillés à des transitions du second ordre. Reproduit d'après l'article de BRYNGELSON.

21. Ou du moins, on s'attend à un ralentissement extrême de la dynamique comme dans le cas d'un verre.

22. Dans la figure, ces paramètres sont tous les deux divisées par la température T afin d'obtenir des quantités adimensionnées

23. R Mélin, H Li, N S Wingreen, and C Tang. Designability, thermodynamic stability, and dynamics in protein folding: a lattice model study. *The Journal of chemical physics*

Que faire d'une protéine repliée ?

Aussi complexe que soit le problème du repliement, ce n'est pas le seul que posent les protéines. L'objectif final étant d'une part d'être capable de déterminer la fonction d'une protéine à partir de sa séquence primaire, d'autre part de concevoir des protéines

capables de remplir une tache précise dans un organisme vivant. Au nombre des étapes qu'il nous reste à franchir pour comprendre l'architecture des protéines, citons donc la réalisation de la fonction, les moyens mis en oeuvre par ces dernières pour résister aux mutations et enfin la façon dont l'on pourrait construire une protéine synthétique *de novo*.

Élucider la fonction

Une fois repliée, chaque protéine doit désormais être capable de remplir sa ou ses fonctions biologiques, fonctions qui, on l'a vu, sont pour le moins variées. Chaque fonction devra sans doute être comprise séparément tant la diversité des rôles joués par la protéine est importante. Comprendre le lien entre la nature des acides aminés et la fonction de la protéine est à la fois intriguant et capital pour être capable de modifier les fonctions de protéines naturelles.

Intriguant car contrairement à ce que laisserait penser une première intuition, même dans le cas de fonctions « localisées » c'est-à-dire réalisées en un site précis de la protéine [24], même si ces derniers sont généralement prépondérants, les résidus les plus proches du sites actifs sont loin d'être les seuls impliqués dans la fonction [25]. Cette hiérarchie peut maintenant être mise en évidence à travers des exemples de mutagenèse saturée et de mesure exhaustive d'activité et révèle une architecture complexe encore largement mal comprise. L'apparition et les contraintes susceptibles de provoquer cette architecture seront justement l'objet de ce tapuscrit.

24. La liaison et l'activité enzymatique par exemple, au contraire l'allostérie qui n'est pas une fonction localisée.

25. R N McLaughlin, Frank J Poelwijk, Arjun Raman, Walraj S Gosal, and Rama Ranganathan. The spatial architecture of protein function and adaptation. *Nature*, 490(7422):138–142, July 2012

Robustesse et Évoluabilité

Ces expériences de mutagenèse mettent en lumière une autre propriété importante des protéines : leur forte propension à demeurer intacte lors d'une mutation. Autrement dit, pour une protéine classique, environ 70% des changements d'acides aminés ne semblent affecter ni le repliement, ni la stabilité, ni la fonction de cette dernière. Cette propriété souvent désignée sous le nom de *robustesse* permet de conserver un organisme fonctionnel malgré les erreurs de copies inévitables [26]. D'un autre côté, elle semble limiter la capacité de l'organisme à s'adapter à un nouvel environnement, une propriété connue sous le nom d'évoluabilité. Nous reviendrons sur ce paradoxe dans le cadre de notre modèle lors de notre conclusion (p. 95).

En plus de ce paradoxe, deux questions se posent naturellement. Tout d'abord, comment l'évolution qui est un phénomène très simple, permet-elle l'émergence de propriété aussi subtil que la robustesse, l'évoluabilité ou tout autres propriétés similaires dont les avantages ne se font sentir qu'au cours de longue période de temps ? Ensuite, on peut se demander ce qui distingue les 30% d'acides aminés dont le changement cause souvent une perte de fonction significative du reste de la protéine ?

26. Claus O Wilke, Jia Lan Wang, Charles Ofria, Richard E Lenski, and Christoph Adami. Evolution of digital organisms at high mutation rates leads to survival of the flattest. *Nature*, 412(6844):331–333, 2001

Conception de protéine

Finalement, l'un des objectifs de l'étude des protéines est d'être capable d'effectuer un lien direct et dans les deux sens entre la séquence d'acides aminés et la fonction. On peut espérer ainsi produire des protéines synthétiques capables d'agir de façon précise dans des régions particulières : venir dégrader de façon spécifique un ensemble d'antigènes dans un organisme, palier une voix métabolique déficiente ou, dans un premier temps, fournir de nouvelles formes de produits chimiques destinés à une utilisation industrielle ou particulière (détergents protéiques par exemple).

Aujourd'hui, ce type de protéine est conçu en utilisant des méthodes inspirées de la sélection « naturelle » [27], d'où l'importance aussi d'une meilleure compréhension et surtout d'une étude quantitative de cette dernière.

27. Sergio G Peisajovich and Dan S Tawfik. Protein engineers turned evolutionists. *Nature methods*, 4(12):991–994, 2007

Pour arriver à produire des protéines destinées à une utilisation médicale, il reste toutefois un problème de taille à résoudre : comprendre et quantifier l'importance de la sélection négative et des interactions parasites. Un organisme vivant comporte en effet, des milliers de protéines, de fragments d'ADN et d'ARN et d'autres molécules rendant extrêmement difficile une action ponctuelle. Pour introduire un nouvel élément, il n'est pas seulement nécéssaire de s'assurer qu'il remplit efficacement son rôle mais aussi qu'il ne perturbe pas excessivement les mécanismes déjà en place dans la cellule. Comprendre et déterminer les principales caractéristiques de ce système hautement frustré sera peut-être le prochain défi de la biologie moléculaire.

Les protèines sont donc des constituants élémentaires du vivant en grande partie définis par leur séquence. Elles possèdent différentes propriétés et remplissent différentes fonctions dont l'on peut chercher les causes dans la répartition des acides aminés à l'aide d'outils physiques et statistiques que nous allons désormais développer.

Les Modèles Gaussiens Élastiques

– Pourquoi les physiciens s'intéressent-ils autant à l'oscillateur harmonique ?
– Parce qu'ils ne savent rien faire d'autre.
Blague répandue dans le milieu des physiciens

La plupart des modèles physiques s'essayant à reproduire des données expérimentales reposent généralement sur de nombreuses hypothèses complexes et un grand nombre de paramètres ajustables. Les Modèles Gaussiens Élastiques (ENM pour *Elastic Network Model* en anglais) forment pour cela un contre exemple intéressant de par leur capacité à retrouver des données complexes à partir d'une hypothèse qui pourrait paraître grossière et de seulement deux paramètres [28].

Nous verrons tout d'abord une présentation du modèle avant d'en discuter les implications dans le cas d'une fonction simple : l'allostérie.

28. On peut par la suite, raffiner le modèle en prenant en compte toujours plus de détails, mais nous montrerons ici un cas où deux paramètres suffisent.

Présentation générale

Le terme de *modèle gaussien* ou *modèle élastique* désigne un ensemble de modèles pouvant paraître simplistes mais ayant de fait un remarquable pouvoir prédictif au vu de leur coût de calcul particulièrement faible. Ils ont été proposés par Tirion [29] puis repris et développés par Bahar [30] entre autres et reposent sur une approximation principale : les mouvements des éléments composant la molécule étudiée restent de faible amplitude et l'on peut donc approximer les potentiels empiriques utilisés ordinairement en dynamique moléculaire par de simples potentiels harmoniques.

Commençons par noter $\vec{x_i}$ les positions des différents éléments de notre système. Une méthode détaillée demanderait de déterminer précisément les potentiels $E_{i,j}(\vec{x_i}, \vec{x_j})$ provenant des différentes interactions (forces de van der Waals, répulsion magnétique, etc.), puis d'en déduire les forces effectives. Deux méthodes sont envisageable :

— Partant de ces forces "fondamentales", on peut effectuer un développement limité afin d'en déduire les termes quadratiques qui seront utilisés lors du calcul,

— On peut aussi supposer ces interactions inconnues et ajuster les différents paramètres afin d'ajuster la courbe expérimentale.

29. Monique M Tirion. Large amplitude elastic motions in proteins from a single-parameter, atomic analysis. *Phys. Rev. Lett.*, 77(9):1905, 1996

30. Ivet Bahar, Ali Rana Atilgan, and Burak Erman. Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. *Folding and Design*, 2(3):173–181, 1997

C'est cette deuxième méthode que nous allons détailler ici. *A priori*, c'est là une chose stupide à faire car le nombre d'atomes dans une protéine est immense et le nombre de paramètres à ajuster rendrait notre méthode ridicule. On peut cependant suivre BAHAR [31] et simplifier le problème jusqu'à obtenir un nombre raisonnable de paramètres.

31. T Haliloglu, I Bahar, and B Erman. Gaussian dynamics of folded proteins. *Phys. Rev. Lett.*, 79(16):3090–3093, 1997

Tout d'abord, supposons connu la conformation stable de la protéine étudiée. On connait donc la position de tous les acides aminés. La seconde approximation consistera à adopter une vision décimée de la protéine dans laquelle on se limitera au mouvement relatif de chaque acide aminé [32]. Le vecteur $\vec{X}$ désigne donc les déplacements des différents acides aminés par rapport à leur position d'équilibre.

32. La question d'utiliser pour coordonnés le centre de masse du résidu ou la position de son carbone *alpha* n'est guère pertinente au vu des approximations drastiques que nous venons déjà de réaliser

Afin de limiter le nombre d'interactions à estimer, la méthode la plus simple est de supposer les interactions de courte portée. On introduit donc une distance de coupure $d_{\text{cut-off}}$ au dela de laquelle les interactions seront simplement ignorées. La valeur de cette distance dépend du niveau de décimation choisie mais est typiquement de $d_{\text{cut-off}} \simeq 10\text{Å}$ pour une représentation des acides-aminés. Nous devrons toutefois vérifier que les résultats ne dépendent pas fortement de ce choix arbitraire.

On s'est donc restreint au hamiltonien d'interaction :

$$\mathcal{H}(\vec{X}) = -\frac{1}{2} \sum_{<i,j>} K_{ij} (\vec{x_i} - \vec{x_j})^2, \tag{5}$$

où $< i, j >$ dénote l'ensemble des couples en interaction, c'est-à-dire tels que $|(\vec{x_i} - \vec{x_j})| < d_{\text{cut-off}}$. On veut désormais en déduire les principaux modes de vibration de la molécule.

Même ainsi, une protéine typique est constituée de 200 acides aminés et comporte 6 fois plus de contacts ce qui représente un nombre de paramètres impressionnant. Notre dernière hypothèse sera donc que la valeur de la constante K_{ij} est identique pour toutes les interactions et tout les résidus. Cette hypothèse très restrictive permet de restreindre à deux le nombre de paramètre libres du modèle : la constante de raideur K et la distance de coupure $d_{\text{cut-off}}$, ce qui est beaucoup plus acceptable.

Nous allons maintenant montrer que, du point de vue mathématique, déterminer les modes de ce hamiltonien revient alors simplement à inverser la matrice des interactions. C'est une matrice de taille N, N étant le nombre d'acides aminés de la protéine étudiées [33]. Cette méthode présente donc l'avantage d'un temps de calcul grandement réduit par rapport à d'autres approches utilisant des énergies potentielles plus complexes et permet d'explorer facilement les déplacements de tout les atomes d'une chaine de taille importante.

33. Nous supposons ici que les couplages entre deux sites ne dépendent que de la distance entre ces sites et sont isotropes.

Retrouver le spectre du facteur B

Le facteur B quantifie le mouvement thermique le long de la chaîne d'acides aminés. Ce mouvement étant dépourvu de direction

privilégiée, ceci conforte notre hypothèse d'isotropie et l'on peut se restreindre à une seule variable par site au lieu de trois.

On a alors une énergie de la forme :

$$\mathcal{H}(\vec{X}) = -\frac{K}{2} \sum_{<i,j>} (\vec{R}_i - \vec{R}_j)^2 = -\frac{K}{2} R^T \Gamma R, \tag{6}$$

où Γ est la matrice laplacienne du graphe de connectivité de la protéine. C'est-à-dire la matrice donnée par [34] :

$$\begin{array}{lcll} \Gamma_{ij} & = & -1 & \text{si } d_{ij} < d_{\text{cut-off}} \\ \Gamma_{ij} & = & 0 & \text{sinon} \\ \Gamma_{ii} & = & \deg(i), & \end{array} \tag{7}$$

où $\deg(i)$ indique le degré, c'est-à-dire le nombre de voisins, du site i.

Pour calculer le facteur B du résidu i, on va simplement regarder la réponse de ce site à une perturbation normée sur ce site : par exemple en ajoutant l'interaction aR_i où a est une constante purement dimensionnelle. La réponse est donnée par :

$$< R_i^2 >= \frac{1}{Z} \int \exp(-\frac{\beta K}{2} R^T \Gamma R + aR_i) dR, \tag{8}$$

ce qui donne $< R_i^2 >= \exp(\frac{a}{2}\Gamma_{ii}^{-1})$ après une simple intégration gaussienne. Autrement dit, le spectre du facteur B est simplement donné par la diagonale de la matrice Γ^{-1}. On pourrait montrer de la même manière que les termes hors-diagonaux donnent les corrélations des fluctuations entre deux sites.

Il est intéressant de noter ici le parallèle entre cette méthode de réseau élastique et l'approximation gaussienne des verres de spins qui a été utilisée pour notre article [35]. La dérivation de l'énergie libre dans le cadre de ce modèle est donnée en appendice (p. 117).

Une telle approche, bien que très simplifiée, permet de retrouver fidèlement les données expérimentales comme on peut le voir dans la figure (6). Dans notre cas, le choix des unités est arbitraire et seules les variations relatives sont significatives, ce qui veut dire que l'on peut fixer $K = 1$. Et on compare les deux courbes à l'aide d'un changement d'échelle.

34. Notons que cette matrice n'est pas nécessairement inversible, elle l'est cependant pour tous les cas étudiés.

35. Mathieu Hemery and Olivier Rivoire. Evolution of sparsity and modularity in a model of protein allostery. *Phys. Rev. E*, 91(4):042704, April 2015

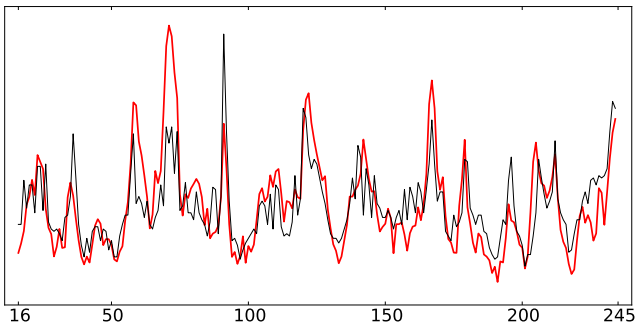

FIGURE 6: Corrélation entre le facteur B expérimental (en rouge) et obtenu par un modèle gaussien (en noir) dans le cas de la Protéase à Sérine du rat (PDB id : 1TON). L'axe horizontal indique la position de l'acide aminé le long de la séquence et l'axe vertical est sans unité. La valeur de cut-off utilisée est $d_{\text{cut-off}} = 8,5$Å. (Calcul personnel)

Le seul paramètre ajustable est alors la distance de coupure $d_{\text{cut-off}}$. On voit que, même si les détails de la structure ne sont pas retrouvés avec précision, on détermine la position des pics et des creux, c'est-à-dire la mobilité générale de la chaîne, avec une bonne fidélité.

Afin de vérifier si le choix de $d_{\text{cut-off}}$ est un critère sensible de notre analyse, nous pouvons tout simplement choisir des valeurs différentes comme le montre la figure (7). Ainsi, du moment que nous choisissons un ordre de grandeur raisonnable, nos calculs permettent d'obtenir des résultats pertinents même à l'aide d'outils relativement simples.

Remarquons pour terminer, que le spectre du facteur B n'est pas la seule quantité que l'on peut déterminer à l'aide d'un tel modèle. En utilisant une description tri-dimensionnelle, on peut retrouver les modes de déplacement principaux de la protéine et accéder ainsi à d'éventuel changements de conformations [36]. Ceci permet aussi de mettre en valeur une séparation de la protéine en blocs rigides correspondant aux différents domaines structurels.

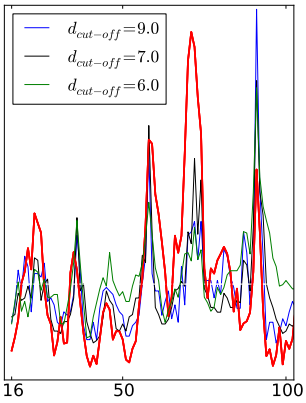

FIGURE 7: Corrélation entre le facteur B expérimental et celui obtenu par un modèle gaussien dans le cas des premiers acides aminés de la Serine Protease pour différentes valeurs de cut-off.

36. Swapnil Mahajan and Yves-Henri Sanejouand. On the relationship between low-frequency normal modes and the large-scale conformational changes of proteins. *Archives of Biochemistry and Biophysics*, 567(C):59–65, February 2015

Cette valeur uniforme de la constante des différents potentiels est particulièrement surprenante car elle laisse à penser que la nature des acides aminées ne controle pas véritablement les modes de vibration et par là, la fonction de la protéine. De fait, la construction d'un modèle gaussien élastique ne fait intervenir que la distribution spatiale des différentes masses constituant la protéine.

Il est cependant à noter que notre méthode ne permet de retrouver que les modes de déplacement globaux qui par définition ne dépendent que peu de la nature exacte des interactions locales et sont donc particulièrement robustes aux variations possibles autour d'un même modèle. Or il n'y a aucune raison de penser que la fonction réelle d'une protéine soit dominée par de tels modes.

Si cette hypothèse semble justifiée dans le cas de fonctions comme l'allostérie ou les moteurs moléculaires [37] qui demande justement une participation de l'ensemble de la protéine souvent à travers un changement de conformation. Elle peut sembler moins vraisemblable pour le cas de fonctions locales telles que la liaison ou l'activité catalytique.

Toutefois, des analyses par des réseaux élastiques peuvent mettre en valeur des fonctions de liaison quand ces dernières demandent une modification de la conformation. De plus l'énergie d'une excitation extérieure de la protéine semble se concentrer fortement au voisinage des sites actifs dans le cas des enzymes [38].

37. Wenjun Zheng and Sebastian Doniach. A Comparative study of motor-protein motions by using a simple elastic-network model. *PNAS*, 100:13253–13258, 2003

38. Francesco Piazza and Yves-Henri Sanejouand. Long-range energy transfer in proteins. *Phys. Biol.*, 6(4):046014, December 2009

Allostérie

On parle d'allostérie lorsque la fonction d'une protéine en un site donnée [39] dépend quantitativement de la présence d'un ligand en un second site. Nous nous restreignons ici au cas d'une allostérie de

39. Constante de liaison, activité catalytique, etc.

liaison : une protéine présentant plusieurs sites de liaison dont les constantes de liaison dépendent de la présence de ligands sur ces autres sites.

L'allostérie est une fonction remarquablement répandue, présente dans un grand nombre de domaines protéiques différents. Elle fut remarquée pour la première fois en 1941 par DISCHE même si le terme d'allostérie n'apparait que vingt ans plus tard sous la plume de MONOD [40]. Un modèle théorique sera rapidement proposé avec l'aide de CHANGEUX [41].

L'exemple le plus connu est sans doute celui de l'hémoglobine qui possède quatre sites actifs, chacun étant d'autant plus actif que les autres sont occupés. Cette configuration permet d'obtenir des réponses (ici capter et libérer l'oxygène) rapides et sensibles.

La vision structurelle proposée alors impliquait que la protéine pouvait adopter plusieurs conformations selon la présence ou non des ligands. Ainsi la liaison du ligand à la protéine (par exemple la liaison du ligand bleu dans la figure 8) modifiait la conformation native (représentée ici par un changement de couleur), modifiant ainsi la liaison du second site (rouge).

La découverte de protéine allostérique non repliée dans leur état natif [42] est venue confirmer les doutes déjà émis sur cette représentation simple [43] et on préférera aujourd'hui parler en terme d'énergie libre de liaison [44], vision que nous adopterons par la suite. En pratique, et si l'on se limite à une vision « à l'équilibre » de l'allostérie, cela signifie que l'on considère non pas seulement l'état natif de la protéine mais l'ensemble des conformations que cette dernière peut adopter.

Une vision structurelle reste cependant utile pour comprendre certains mécanismes d'allostérie, en particulier quand plusieurs conformations stables ont été découvertes pour une même protéine. On peut alors comparer les tensions ou les frustrations de ces différentes structures pour déterminer comment fonctionne la partie énergétique de l'allostérie [45].

La liaison d'un ligand à une protéine peut être quantifié par la différence d'énergie libre entre l'état libre et l'état lié, aussi appelée énergie libre de liaison :

$$\Delta F(\ell_1) = F(\ell_1 = 1) - F(\ell_1 = 0), \tag{9}$$

où $\ell_1 = 1$ et $\ell_1 = 0$ indique respectivement la présence et l'absence du ligand.

L'allostérie est alors donnée par la différence d'énergie libre de liaison en présence et en absence du second ligand [46] :

$$\begin{aligned} \Delta\Delta F(\ell_1, \ell_2) &= \Delta F(\ell_1, \ell_2 = 1) - \Delta F(\ell_1, \ell_2 = 0), \\ &= F(1,1) - F(0,1) - F(1,0) + F(0,0). \end{aligned} \tag{10}$$

Cette énergie libre peut être calculée de nombreuses manières, notamment à l'aide d'expérience de calorimétrie par exemple. Mais

40. G N Cohen. Regulation of enzyme activity in microorganisms. *Annual Reviews in Microbiology*, 1965

41. Jacques Monod, Jeffries Wyman, and Jean-Pierre Changeux. On the nature of allosteric transitions: a plausible model. *Journal of Molecular Biology*, 12(1-2):88–118, 1965

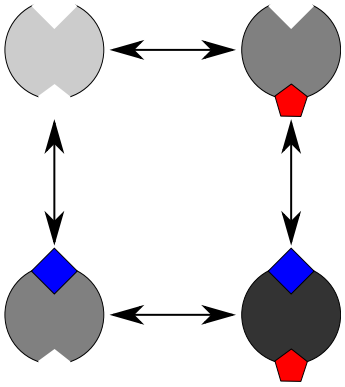

FIGURE 8: Schéma simplifié d'allostérie de liaison à deux sites.

42. Abel Garcia-Pino, Sreeram Balasubramanian, Lode Wyns, Ehud Gazit, Henri De Greve, Roy D Magnuson, Daniel Charlier, Nico A J van Nuland, and Remy Loris. Allostery and Intrinsic Disorder Mediate Transcription Regulation by Conditional Cooperativity. *Cell*, 142(1):101–111, July 2010

43. Rhoda J Hawkins and Tom C B McLeish. Coarse-Grained Model Of Entropic Allostery. *Phys. Rev. Lett.*, 93(9):098104, 2004

44. Hesam N Motlagh, James O Wrabl, Jing Li, and Vincent J Hilser. The ensemble nature of allostery. *Nature*, 508(7496):331–339, 2014

45. D U Ferreiro, J A Hegler, and E A Komives. On the role of frustration in the energy landscapes of allosteric proteins. *PNAS*, 2011

46. Notons que même si les deux sites sont strictement équivalents, lorsque la fonction peut être clairement identifiée, on préférera parler de modulateur ou de régulateur pour désigner le second ligand.

il est intéressant de disposer d'un modèle prédictif pour quantifier l'allostérie d'une protéine sans recourir à des expériences.

Dans ce cadre, une application intéressante des modèles élastiques proposée par Rodgers *et al.*[47] est d'utiliser cette approximation pour calculer l'énergie libre des différents états de la protéine et d'en déduire une valeur théorique de l'allostérie.

Mais plus intéressant encore de notre point de vue, on peut modéliser une protéine allostérique à l'aide d'un modèle élastique puis observer l'effet d'une modification de la constante d'un couplage particulier afin de mettre en lumière les contacts ayant la plus forte influence sur l'activité allostérique $\Delta\Delta F$[48], une manière de chercher les couplages les plus essentiels.

Cette analyse permet une compréhension fine, avec l'aide d'un modèle très simple, de la façon dont une protéine réalise sa fonction et répond à une mutation. Elle permet donc d'une part de connaître les acides aminés les plus impliqués dans la fonction, d'autre part de sonder la géométrie fonctionnelle d'une protéine, c'est-à-dire les régions de cette dernière ou se concentrent les couplages essentiels à la fonction. On pourra alors comparer cette géométrie à celle résultant d'expériences de mutation ou d'analyse de séquence que nous définirons dans le chapitre consacré aux secteurs (p. 51).

47. Thomas L Rodgers, David Burnell, Phil D Townsend, Ehmke Pohl, Martin J Cann, Mark R Wilson, and Tom CB McLeish. ΔΔ PT: a comprehensive toolbox for the analysis of protein motion. *BMC Bioinformatics*, 14(1):183, 2013

48. Thomas L Rodgers, Philip D Townsend, David Burnell, Matthew L Jones, Shane A Richards, Tom C B McLeish, Ehmke Pohl, Mark R Wilson, and Martin J Cann. Modulation of Global Low-Frequency Motions Underlies Allosteric Regulation: Demonstration in CRP/FNR Family Transcription Factors. *PLoS Biol*, 11(9):e1001651, September 2013

Les modèles gaussiens mettent en évidence une propriété remarquable des protéines. De nombreuses propriétés ne dépendent pas (à l'ordre le plus grossier) de la nature des différents acides aminés. Pourtant, une fonction comme l'allostérie semble se concentrer sur l'utilisation d'un nombre restreint de couplage et suivre une sorte de chemin pour porter son « message » d'un bout à l'autre de la protéine. Est-ce à dire que certains sites sont plus important que d'autre ? Lesquels, pourquoi et comment les trouver sera l'objet de la suite de ce tapuscrit. Nous n'avons cependant que trop tarder pour introduire l'élément central de notre étude : l'évolution.

Évolution naturelle et algorithmique

> Given the presence of certain matter with specified autocatalytic properties and under the maintenance of the finite (free) energy flow necessary to compensate for the steady production of entropy, evolution appears to be an inevitable event.
>
> Manfred Eigen, *Die Naturwissenschaften*, 1971

La théorie de l'évolution est un concept souvent mal compris du grand public, ce qui peut poser de nombreux problèmes lorsque l'on veut en discuter les tenants et les aboutissants. Aussi nous parait-il utile, avant d'entrer dans le vif du sujet, d'en rappeler les grandes lignes historiques ainsi que de donner un bilan de la théorie actuelle qui reste encore sujette à débat.

Nous en profiterons pour présenter un domaine de l'informatique qui lui est intimement lié : celui des algorithmes génétiques.

Un peu d'histoire

Il est toujours difficile de remonter aux premiers pionniers d'une théorie même si, dans le cas de la théorie de l'évolution, il est certain que Lamarck et Darwin furent ceux qui en permirent l'essor durant la seconde partie du XIXème siècle. Leurs idées se diffusèrent rapidement au sein de la communauté scientifique mais c'est dans les détails de ces dernières que se cachaient les nombreuses controverses qui animèrent la discipline durant cinq décennies. Débats fortement nourri par la personnalité des chefs de file, il faut le reconnaître [49].

Que les individus varient et donnent naissance à de nouvelles espèces par un processus d'hérédité et de mutations fut immédiatement tenu pour donné. Non seulement car ces concepts étaient déjà vivaces depuis quelques temps, mais aussi car la somme de preuves fournie par Darwin dans son maître ouvrage *De l'origine des espèces* [50] ne laissait guère de place au doute. Ce sont les nombreux espaces qu'il restait à remplir qui furent l'objet d'âpres débats.

La nature du vecteur de l'information héréditaire en est un exemple. Comment, où et quand est écrit le plan qui déterminera la nature des descendants d'un individu ? Est-il créé une fois pour toutes lors de la naissance ou bien tout au long de la vie en agrégeant l'information recueillie par l'individu ? La question est importante car elle détermine si des modifications ou mutations survenant au cours de la vie d'un individu peuvent se transmettre

49. W. B. Provine. *The Origins of Theoretical Population Genetics*. The University of Chicago Press, 1971

50. C. Darwin. *On the origin of species by means of natural selection*. J. Murray, 1859

à sa descendance. Est-ce en tendant leurs cous pour atteindre les feuilles que les girafes l'ont allongé ?

On sait maintenant grâce à la découverte de l'ADN que le développement d'un individu n'influe pas sur le patrimoine génétique qu'il léguera à ses descendants… encore que [51].

Le processus menant à l'apparition de nouvelles espèces fut lui aussi très discuté. Ces dernières apparaissent-elles soudainement avec la venue au monde d'un mutant présentant une forte variation (on pense à une nouvelle paire d'ailes sur une mouche, ou une nouvelle couleur chez une fleur) comme le proposait Galton ou bien de manière progressive en modifiant petit à petit un caractère déjà existant (allongement de la longueur des pattes, par exemple) comme le soutenait Pearson ?

Il fallut attendre la redécouverte des travaux de Mendel et leur fusion avec les idées de l'évolution pour voir esquisser les contours de la théorie synthétique de l'évolution qui reste encore aujourd'hui le paradigme principal de la biologie moderne. Même si la découverte de l'ADN et la riche période qui s'ensuivit devait encore en modifier quelque peu la structure.

51. Rodolphe Barrangou, Christophe Fremaux, Hélène Deveau, Melissa Richards, Patrick Boyaval, Sylvain Moineau, Dennis A Romero, and Philippe Horvath. CRISPR provides acquired resistance against viruses in prokaryotes. *Science*, 315(5819):1709–1712, 2007

La théorie synthétique de l'évolution par sélection

La théorie de l'évolution, bien que conçue dans le cadre de la biologie, n'y est pas nécessairement limitée. Pour cela il suffit qu'une classe d'objets [52] regroupe les trois caractères de copie, de compétition et de variation, il s'en suivra naturellement une forme ou une autre d'évolution comme le fait remarquer la citation de Eigen au début de ce chapitre.

la Copie : qui permet de créer une nouvelle instance à partir de celles déjà existantes. Dans le cas où ces instances ont une durée de vie finie, ce mécanisme permet de plus d'assurer une certaine forme de mémoire non locale.

la Sélection : ou toute autre forme de compétition ou de comparaison – entre les instances qui permet de retirer les propositions infructueuses et de se concentrer sur celles qui ont prouvées leur efficacité. Déterminer quel est l'objectif sous-jacent peut être très facile – dans un algorithme génétique ce dernier est défini par l'utilisateur – ou très complexe, comme c'est le cas en biologie.

les Variations : par l'introduction de différences lors de la copie, différences qui doivent pouvoir jouer un rôle lors de la compétition. L'objectif étant d'explorer au mieux l'ensemble des solutions possibles, en l'absence de variation, on ne pourra au mieux que déterminer la meilleure des solutions proposées initialement.

Deux points très importants méritent d'être signalés.

Premièrement, en l'absence de sélection, le système continuera à évoluer, c'est-à-dire à se modifier dans le temps, mais de manière purement aléatoire. C'est le cas où bien quand la pression

52. Le vocabulaire de la programmation est particulièrement adaptée pour cette définition : une *classe* donne la structure et la définition d'*objets* partageant certaines propriétés ; une *instance* désigne l'un de ces objets en particulier, c'est donc une réalisation de la classe. Je suis une instance de la classe des systèmes biologiques.

de sélection est si faible que même des individus très peu adaptés peuvent survivre et se reproduire ou bien quand les variations introduite ne jouent aucun rôle vis à vis de la sélection. On parlera alors d'*évolution neutre*, le rôle de cette dernière par rapport à l'*évolution sélective* étant d'un intérêt crucial et soumis à d'ardent débats. Contentons nous donc de garder à l'esprit que la *sélection naturelle* est l'une des forces principales de l'*évolution naturelle* mais que cette dernière ne se réduit pas à la sélection et que certaines caractéristiques complexes peuvent être le produit de l'évolution neutre [53].

Secondement, le mécanisme de sélection fonctionne avec un temps caractéristique donné. Si l'information se dégrade d'elle même plus rapidement, par exemple si les différentes copies se modifient ou si les copies disparaissent et que le mécanisme de copie génèrent trop de variation, le système est gouverné par la stochasticité, c'est le régime dit du seuil d'erreur (*error threshold* en anglais).

53. Si vous êtes surpris, pensez à une marche aléatoire sur un graphe non régulier, les noeuds de haut degré seront naturellement visités plus souvent par ce processus 'neutre'.

Nous l'avons signalé, la biologie n'est pas le seul domaine où se présente une forme de sélection et d'évolution. Même si c'est sur cette dernière que nous nous concentrerons par la suite, nous donnons ici brièvement quelques exemples de sélection différentes. Notez que chacun de ces domaines représente un cas particulier avec ses propres lois et dynamiques :

en Biologie les individus sont encodés dans l'ADN qui permet à la fois l'hérédité et une forme de variation et sont soumis à une lutte pour la survie et la reproduction induisant une sélection entre espèces partageant un même territoire au sens le plus large du terme.

en Économie les entreprises cherchent à établir des stratégies leur permettant d'être rentables et de s'agrandir. La copie et la variation proviennent principalement des agents humains qui composent ces "organismes".

en terme de Culture les innovations technologiques et les cultures évoluent au fil du temps, au fur et à mesure des découvertes et des modes. Certaines cultures progressent plus rapidement car elles permettent à ceux qui les partagent de se développer plus rapidement, ou bien car elles bénéficient d'un certain attrait amenant d'autre groupe à les intégrer. La question de savoir si ce domaine est effectivement une forme de sélection est ardemment débattue.

en Science les théories scientifiques suivent une évolution en cherchant une meilleure description de la réalité, la compétition reposant généralement sur le pouvoir prédictif, la mémoire provenant aujourd'hui principalement d'outils technologiques (livres, serveurs, etc.) et une variation dirigée étant fourni par les scientifiques.

Pour autant l'évolution biologique, c'est-à-dire la modification dans le temps des organismes vivants, présente certaines propriétés rendant son étude d'autant plus complexe et plus intéressante. Commençons par examiner comment les trois caractères fondamentaux y sont représentés. Puis nous passerons en revue certaines de ses particularités.

La *copie* est assurée au niveau d'une population par la reproduction de ses différents individus. Le principal support de l'hérédité est évidemment l'ADN qui est hérité entre les organismes vivants, généralement du parent vers le descendant, mais pas seulement [54]. Le mécanisme de copie de l'ADN est remarquablement fidèle avec un taux d'erreur (c'est-à-dire le ratio d'erreurs d'une génération à la suivante) variable selon les organismes allant de 10^{-4} chez les virus à 10^{-11} chez les bactéries [55]. D'autre part, de nombreuses caractéristiques en dehors de l'ADN peuvent favoriser un individu par rapport à un autre : sa position géographique, la culture dans laquelle il est né, les molécules transmises par ses aïeuls, etc. Tous ces facteurs sont regroupés sous la notion d'*épigénétique* et forment aujourd'hui un champ de recherche particulièrement actif tant du point de vue théorique que pratique.

La *sélection* est assurée par la capacité d'un individu à se reproduire de façon efficace, c'est-à-dire à permettre à son patrimoine génétique d'être présent au sein des populations futures, que ce soit en ayant davantage de descendants, en s'assurant que ces derniers survivent jusqu'à l'âge adulte ou tout autre méthode favorisant directement ou indirectement ses propres gènes. Notez que "mes gènes" ne signifient pas toujours forcément "moi-même" : il suffit d'avoir à l'esprit les insectes sociaux comme les fourmis qui possèdent des gènes très semblables au sein d'une même colonie pour comprendre cette différence.

Cette capacité à passer ses gènes dans la génération suivante est souvent désignée par le terme de "fitness", mais sa définition est sujette à d'interminables débats dans lesquels nous préférerons ne pas entrer dans le corps de cet ouvrage, préférant lui consacrer un appendice (p. 113).

La *variation* enfin dépend de l'organisme dont il est question. Chez les bactéries il s'agit principalement d'erreur de copie lors de la réplication de l'ADN mais aussi de transport d'un gène d'une partie du génome à une autre et de transferts horizontaux. Chez les espèces sexuées, la variation provient aussi de l'arrangement des gènes entre paires et des nombreux cross-over ainsi que des transposons [56], la mutation d'un seul site n'étant plus la principale source de variation.

Ces variations ne sont donc pas uniquement aléatoires comme on le croit bien souvent. Les bactéries peuvent jouer avec leur taux de mutation [57] et échanger "volontairement" des groupes de gènes pour répondre à un environnement hostile.

54. Les Transferts de Gènes Horizontaux (HGT) présents chez de nombreuses bactéries, permettent des passages de matériel génétique entre deux individus quelconques, y compris entre des espèces distinctes.

55. À titre de comparaison, ce taux est de 10^{-12} sur un disque dur actuel, avant l'utilisation de codes correcteurs.

56. Les transposons sont des séquences d'ADN capable de se déplacer – mais pas de se répliquer – de manière autonome dans un génome.

57. Ivana Bjedov, Olivier Tenaillon, Benedicte Gerard, Valeria Souza, Erick Denamur, Miroslav Radman, François Taddei, and Ivan Matic. Stress-induced mutagenesis in bacteria. *Science*, 300(5624):1404–1409, 2003

De plus, les organismes vivants sont des entités auto-répliquantes, ce qui implique que les mécanismes par lesquelles une espèce évolue sont eux-mêmes soumis à la contrainte de l'évolution [58]. Cette particularité est loin d'être anecdotique, elle implique que la matière biologique ne suit pas des règles fixes – hormis les lois de la physique – mais fait évoluer ses propres règles. Le taux de mutation, le code génetique [59] ou encore les mécanismes de réarrangement devraient donc pouvoir être étudiés, du moins en partie, à l'aide à la théorie de l'évolution.

58. O Rivoire and S Leibler. A model for the generation and transmission of variations in evolution. *PNAS*, 111(19):E1940–E1949, May 2014

59. Kalin Vetsigian, Carl Woese, and Nigel Goldenfeld. Collective evolution and the genetic code. *arXiv*, q-bio.PE, May 2006

Biologie et évolution

Pour autant qu'elle soit désormais comprise, la théorie de l'évolution n'est pas encore pleinement exploitée au sein de la biologie. En effet, la façon dont l'évolution échantillonne les différentes solutions possibles d'un même problème est loin de ressembler à la façon dont un ingénieur chercherait ces dernières [60]. Il est donc possible, en observant un système, de déterminer s'il est le fruit d'une évolution ou non. Une montre suisse trouvée sur le chemin trahit la main d'un ouvrier conscient de son but tandis que le génome d'une bactérie laisse entrevoir le fruit de l'évolution naturelle. Autrement dit, le mécanisme qui a fourni la solution doit laisser sa trace dans l'architecture de cette dernière.

60. François Jacob. Evolution and tinkering. *Science*, 196(4295):1161–1166, 1977

L'étude détaillée des séquences génétiques devrait donc nous permettre de comprendre plus finement le détail de l'évolution naturelle mais aussi nous apporter des indices sur l'histoire de la pression de sélection qui fut appliquée aux organismes dont nous contemplons aujourd'hui les descendants. L'évolution par sélection peut en effet être pensée comme un processus d'apprentissage [61] au cours duquel la population s'adapte progressivement à l'environnement dans lequel elle vit et évolue. En determinant les caractères qui ont été sélectionnés et la manière dont ils l'ont été, on pourrait tirer de nombreuses conclusions sur le passé de notre planète.

61. Olivier Rivoire and Stanislas Leibler. The Value of Information for Populations in Varying Environments. *J Stat Phys*, 142(6):1124–1166, March 2011

Pour autant, il ne faut pas succomber à la tentation d'attribuer à toute observation une hypothèse évolutive [62]. Par nature, le processus de l'évolution repose sur des tentatives grandement aléatoires pour trouver des solutions et certaines observations pourraient bien être dues à ce hasard plutôt qu'à une quelconque sélection. L'importance respective de ces deux forces étant encore au cœur des questions de recherches actuelles.

62. S J Gould and R C Lewontin. The spandrels of San Marco and the Panglossian paradigm: a critique of the adaptationist programme. *Proceedings of the Royal Society B: Biological Sciences*, 205(1161):581–598, 1979

Cependant une meilleure compréhension du fonctionnement de l'évolution pourrait nous amener à de meilleures méthodes de conception des protéines de synthèse ou des médicaments, par exemple en anticipant les prochaines souches de virus [63]. Aujourd'hui, l'approche privilégiée est celle de la force brute : on identifie tout d'abord des solutions naturellement proches de la fonction recherchée puis on impose à la population une succession de pressions de sélection et de mutations pour faire émerger de nouvelles solutions [64]. Il est cependant possible que l'histoire de la pression

63. Richard A Neher, Colin A Russell, and Boris I Shraiman. Predicting evolution from the shape of genealogical trees. *eLife*, 3, November 2014

64. Sergio G Peisajovich and Dan S Tawfik. Protein engineers turned evolutionists. *Nature methods*, 4(12):991–994, 2007

de sélection ait une grande influence [65], de même que le choix de la composition de la population initiale ; comprendre le rôle et le fonctionnement de ces éléments parait donc particulièrement important.

Une biologie prédictive

La question de la prédiction en biologie est une question ancienne est particulièrement débattue [66] : Est il possible de prédire qualitativement ou quantitativement l'évolution d'un système biologique de la même manière que la description d'un système physique permet de déterminer son état à des temps ultérieurs ?

À première vue, la vie semble travailler dans un régime drastique de sous échantillonnage. C'est-à-dire que l'ensemble des solutions possibles est si vaste que les trajectoires des systèmes sont invariablement dominées par les effets de tailles finies si bien qu'il semble impossible de faire une quelconque prédiction. Plusieurs arguments, théoriques [67] et expérimentaux viennent toutefois perturber cette première idée : tout d'abord, les contraintes pesant sur les séquences sont elles aussi nombreuses et restreignent d'autant plus les chemins qu'il est possible à l'évolution de suivre. La balance entre les degrés de liberté et les contraintes reste cependant loin d'être déterminée. De plus, des expériences récentes d'écosystèmes clos avec plusieurs espèces montrent une forme de reproductibilité de l'évolution qui pourrait être un premier indice d'une possible biologie prédictive, même à l'échelle d'une population [68].

Dans tout les cas, développer des outils permettant une étude quantitative des différents paramètres semble être une première nécessité.

65. Jesse D Bloom, Claus O Wilke, Frances H Arnold, and Christoph Adami. Stability and the evolvability of function in a model protein. *Biophysical Journal*, 86(5):2758–2764, 2004

66. S.J. Gould. *Wonderful Life: The Burgess Shale and the Nature of History*. New York: W.W. Norton & Co., 1989

67. Ivan G Szendro, Jasper Franke, J Arjan GM de Visser, and Joachim Krug. Predictability of evolution depends nonmonotonically on population size. *PNAS*, (2):571–576, 2013

68. Doeke R Hekstra and Stanislas Leibler. Contingency and Statistical Laws in Replicate Microbial Closed Ecosystems. *Cell*, 149(5):1164–1173, May 2012

Algorithmes génétiques

C'est justement pour comprendre la façon dont fonctionne l'évolution qu'Holland a développé une branche importante de l'informatique aujourd'hui connu sous le nom d'algorithme génétique. Une telle démarche n'était certes pas nouvelle comme en témoigne les travaux pionniers de Von Neumann et surtout Barricelli, mais c'est indubitablement lui qui en permit l'essor et proposa un cadre théorique pour décrire et comprendre le processus d'évolution dans le vivant.

L'objectif de Holland et de ses successeurs a toujours été double, cherchant d'une part à utiliser la puissance de l'évolution pour l'appliquer à des problèmes complexes pour lesquels il n'existait pas de méthodes efficaces de résolution, d'autre part à comprendre à la fois qualitativement et quantitativement cette dernière.

Avant de nous lancer dans les notations qui nous serviront par la suite, il est cependant nécéssaire de spécifier la manière dont fonctionne l'évolution et de souligner les principaux aspects des algorithmes génétiques.

En effet, contrairement à la plupart des algorithmes, l'évolution effectue un calcul fortement parallélisé avec une mémoire dispersée et un très fort sous-échantillonnage de l'espace des solutions afin de chercher non pas la meilleure solution, mais *une trajectoire évolutive maximisant la solution à tout instant*. Un algorithme génétique permet ainsi de trouver une "bonne" solution en un temps court par rapport à d'autres méthodes. Évidemment, une telle solution n'est pas idéale pour tout les problèmes [69], mais on retrouve aujourd'hui de tels algorithmes dans un grand nombre de domaines différents : traitement d'image, optimisation de fonction de nombreuses variables, contrôle de systèmes industriels, repliement de protéines, etc. Ces derniers disposent en effet d'une bonne tolérance au bruit [70] et sont, par nature, aisément parallélisables ce qui permet de réduire grandement le temps de calcul pour peu que l'on en ait les moyens.

Le principal défaut provient du manque de compréhension théorique du fonctionnement de ces algorithmes ce qui rend leur utilisation délicate dans certaines situations et complique le choix de l'algorithme à utiliser lors de l'implémentation pratique d'un problème [71].

69. Melanie Mitchell, John H Holland, and Stephanie Forrest. When will a genetic algorithm outperform hill climbing? In J D Cowan, G Tesauro, and J Alspector, editors, *NIPS*, pages 51–58. Advances in Neural Information Processing Systems, 1994

70. B L Miller and D E Goldberg. Genetic algorithms, selection schemes, and the varying effects of noise. *Evolutionary Computation*, 1996

71. Q Pham. Competitive evolution: a natural approach to operator selection. *Progress in evolutionary computation*, pages 49–60, 1995

Notations

Afin de discuter efficacement de l'influence des différents paramètres dans le processus d'évolution et de mettre en évidence la structure commune des différents modèles utilisés., nous utiliserons les notations suivantes – fortement influencées par *Adaptation in natural and artificial systems* [72] :

72. J.H. Holland. *Adaptation in natural and artificial systems*. The MIT Press, 1992

— Le *génotype* ou génome noté G lorsqu'il s'agit d'un génome particulier et $\mathcal{G}$ pour l'ensemble des génomes possibles

— Le *phénotype* noté P lorsqu'il s'agit d'un individu particulier et $\mathcal{P}$ pour l'ensemble des phénotypes possibles

— Le *plan de développement* noté d à valeur dans $\mathcal{P}$, qui associe le phénotype au génotype. Il peut *a priori* dépendre de l'environnement. $d(E, G)$

— La *fécondité* notée ϕ est une variable aléatoire entière quantifiant le nombre de descendants d'un individu (c'est-à-dire d'un phénotype donné) dans un environnement. À chaque environnement est ainsi associée une densité de probabilité sur $\mathbb{N}$, $\phi(E, P)$

— La *sélection* noté σ est une variable aléatoire booléenne quantifiant la probabilité que l'individu survive à la période en cours, c'est une fonction de l'environnement et du phénotype : $\sigma(E, P)$

— Le *plan d'évolution* noté μ à valeur dans $\mathcal{G}$, qui détermine le génotype des descendant d'un individu donné. $\mu(G)$ ou $\mu(G_1, G_2)$ dans le cas d'une reproduction sexuée [73].

Chacune des fonctions décrites précédemment (la fécondité notamment, mais aussi les plans d'évolution et de développement)

73. On suppose généralement que cette fonction ne dépend que du génotype, il est cependant possible que le phénotype ou l'environnement interviennent comme argument de μ, on parlera alors de *lamarckisme*.

sont *a priori* des variables aléatoires. La fécondité d'un individu i peut donc désigner $\phi(E, P_i)$ la variable aléatoire à valeur dans $\mathbb{N}$ déterminant le nombre d'enfant de ce dernier ou ϕ_i une réalisation de la variable précédente. On utilisera alors la notation $\phi_i \overset{t}{=} \phi(E, P_i)$ pour indiquer que le premier terme est une réalisation du second.

Notez également que le terme d'environnement contient, notamment lors du calcul de la sélection, le reste de la population. C'est par l'intermédiaire de ce terme qu'apparaît la compétition entre les différents individus au sein de la population.

Le processus complet de sélection naturelle peut alors être décrit de la manière suivante.

On part d'une population $\mathbb{P}(t) = \{(G_i, P_i)\}$ où l'indice i parcourt les différents individus. On évalue alors la fécondité $\phi_i \overset{t}{=} \phi(E, P_i)$ et la sélection $\sigma_i \overset{t}{=} \sigma(E, P_i)$ de chaque individu ce qui nous permet alors de construire la génération à l'étape suivante[74] :

$$\mathbb{P}(t+1) = \sum_i \sigma_i \times \{(G_i, P_i)\} + \phi_i \times \{(\mu(G_i), d_E(\mu(G_i)))\}, \qquad (11)$$

74. Avec les conventions naturelles $\{A\} + \{B\} = \{A, B\}$ et $2 \times \{A\} = \{A, A\}$.

Le premier terme de la somme indique si l'individu survit à la période, le second terme ajoute ses enfants à la population. Notez que l'on suppose implicitement que les individus se développent dans l'environnement qui les a vu naître[75].

75. On pourrait aussi imaginer inclure un terme de vieillissement des individus : $P_i(t+1) = v(E, P_i)$.

Selon les cas, on peut ensuite terminer le processus lorsque la population atteint une forme d'équilibre, un score jugé suffisant ou au bout d'un temps pré-défini.

Le processus décrit ci dessus est un processus en temps discret, mais son formalisme permet de passer en temps continu très aisément. Il suffit de rendre l'intervalle de temps petit $\mathbb{P}(t) \to \mathbb{P}(t+dt)$ avec les fonctions $\phi, et (1-\sigma)$ d'ordre au moins un en dt.

Pour le choix du lien entre génotype et fécondité, de nombreuses méthodes ont été développées et chacune dispose d'avantages et d'inconvénients. Essentiellement, on désire une fonction associant à un génotype donné une variable aléatoire de $\mathbb{N}$ telle que la taille de notre population soit constante, et qu'une forme de relation d'ordre de $\mathcal{G}$ passe à l'espérance de notre variable[76]. Si possible, on souhaiterait que la variance de cette variable au sein de la population soit toujours de l'ordre de $\frac{1}{2}$. Cette dernière condition est garante d'une sélection efficace car c'est la différence entre le nombre d'enfants des différents individus dans la population qui permet à l'évolution d'effectuer son travail de sélection tout en garantissant une certaine mémoire de la population.

76. Bref, on veut qu'un « meilleur » génotype ait en moyenne plus d'enfants.

Cependant, ce critère est très difficile à remplir si l'on souhaite utiliser une fonction qui ne dépende que de G. En effet, au début de la simulation les différences entre génotypes sont importantes, trouver une amélioration significative est plutôt courant tandis qu'à la fin de la simulation la population est globalement homogène et les mutations permettant un effet bénéfique sont rares. Une telle

fonction ne permet pas d'ajuster le nombre d'enfants au fur et à mesure des progrès de l'évolution, voici deux solutions couramment utilisées[77].

77. M. Mitchell. *An Introduction to Genetic Algorithms*. The MIT Press, 1998

On peut choisir de noter chaque individu selon un seul nombre réel, le score s, puis d'effectuer un *sigma-scaling* c'est à dire la transformation suivante :

$$\varphi(s) = 1 + \frac{s_i - \bar{s}}{2\sqrt{\sigma_s}}, \tag{12}$$

où l'on a notés $\bar{s}$ & σ_s respectivement la moyenne et la variance de s au sein de la population. $\varphi(s)$ dispose de nombreuses propriétés intéressantes, c'est en effet une variable réelle dont la moyenne au sein de la population vaut 1, ce qui assure une taille constante et dont la variance vaut $\frac{1}{4}$ ce qui assure pour des distributions raisonnables que $\varphi(s)$ fluctuera entre 0 et 2 et permettra une évolution efficace.

On introduit ensuite la variable aléatoire $V(x)$:

$$\begin{array}{rcll} V(x) & = & E(x) & \text{avec probabilité } 1 - (x - E(x)), \\ & = & E(x) + 1 & \text{avec probabilité } x - E(x), \end{array} \tag{13}$$

où $E(x)$ dénote la partie entière de x. Ceci permet de palier au fait que $\varphi(s)$ est un nombre réel : on utilise $V(\varphi(s))$ comme fécondité. Cette méthode possède l'avantage de garder une notion quantitative du score, les différences de score se retrouveront réellement dans le nombre d'enfants des différents individus. Le principal problème viendra des cas où il existe une grande probabilité de mutations fortement délétères, on a alors un effet de contraste parasite qui affectera une fécondité fortement négative à certains individus et nécessite de nouveaux traitements *ad-hoc* pour éviter que cet artefact ne bloque l'évolution du reste de la population.

Pour éviter ce type de problème, la *stratégie élite* propose un choix radical. Après avoir trié la population, on affecte un enfant au f% meilleurs de la population[78] et on supprime les f% moins bons (fécondité nulle). Cette méthode ne conserve pas la notion quantitative du score mais permet d'éviter des aberrations telle que celle décrite précédemment et permet de ne pas recourir à des nombres aléatoires ce qui est un réel avantage en terme de temps de calcul aussi bien que pour d'éventuelles analyses théoriques. La taille de la population et la variance de la fécondité sont aussi toutes deux constantes. Cette stratégie est intéressante aussi car elle ne necessite qu'un ordre relatif sur l'ensemble des génotypes, ceci élimine donc un grand nombre de choix arbitraire quand à la gradation de la fécondité.

78. Des valeurs typique de f allant de 10% à 50% et permettent de moduler l'intensité de la pression de sélection.

Autres choix particulièrement cruciaux, le génotype et le plan d'évolution. Nous choisissons d'en parler simultanément car ils sont intimement mêlés et gouvernent une grande partie de la dynamique du modèle. En effet, la façon dont les différents individus sont décrits (le génotype) contraint de manière évidente la façon dont ces éléments seront modifiés (le plan d'évolution). Or une part

importante du succès d'un algorithme génétique réside dans la capacité du plan d'évolution à produire des solutions viables et plus précisément, à mélanger/modifier les solutions viables afin d'en produire de plus efficaces [79].

79. S. Nolfi and D. Floreano. *Evolutionary Robotics*. The MIT Press, 2000

Cas limites et cas pratiques

Cas des plantes annuelles

Il est courant (et nous le ferons majoritairement par la suite) d'utiliser une approximation où la population est entièrement renouvellée à chaque pas de temps. Il suffit alors de prendre $\sigma = 0$. Cette description a le mérite de fixer une échelle naturelle de temps : la génération. Comme la plupart de nos modèles suivront cette approximation, nos résultats seront majoritairement exprimés en génération (l'environement fluctue avec une période de 150 générations, etc.). Même si cette approximation semble manquer de "réalisme" – encore que l'on pourrait imaginer décrire une population de plantes annuelles – elle permet d'accélérer considérablement les simulations et rend caduque la fonction σ, supprimant ainsi une hypothèse et réduisant d'autant les paramètres nécéssaires pour décrire l'évolution.

Taille de population infinie

Une limite intéressante à mentionner d'un point de vue conceptuel est le cas d'une population de taille infinie. En effet, le plan d'évolution permet de passer de n'importe quel génotype à n'importe quel autre fut-ce dans des probabilités faibles ou dans des temps suffisamment long. Dans le cas d'une population infinie cela implique donc qu'au bout d'un régime transitoire toutes les possibilités sont testées simultanément et qu'il existe en permanence au moins un représentant de chaque génotype au sein de la population. Bien sûr, c'est le phénotype ayant la meilleure croissance qui domine (exponentiellement) la population.

Cependant, en cas de changement d'environnement, le génotype le plus adapté peut changer, mais alors que l'on pourrait penser que le plan d'évolution est essentiel pour trouver le nouveau maximum, il est tout à fait accessoire. En effet, *le génotype le mieux adapté est toujours déjà présent dans la population* ; il se contente de prendre le pas exponentiellement rapidement sur le précédent.

Le cas limite d'une population infinie est donc tout à fait décevant sur le plan expérimental car il ne correspond jamais à un cas pratique. Mais il souligne efficacement le rôle des mutations dans l'évolution : assurer qu'il n'y ait pas de génotype (ou d'allèle) qui disparaisse définitivement de la population [80].

80. J.H. Holland. *Adaptation in natural and artificial systems*. The MIT Press, 1992

Sélection forte, Mutation faible

Un régime couramment utilisé dans les travaux théoriques sur l'évolution est celui dit de Sélection forte, Mutation faible [81] (SSWM, Strong selection, Weak Mutation en anglais). Dans ce modèle, on suppose que le taux de mutation est suffisamment faible pour que la population ne contienne qu'un seul génotype, tandis que la sélection est suffisamment forte pour empêcher l'apparition de mutation neutre. Toute mutation est donc bénéfique ou délétère.

La situation se rapproche alors de celle d'un algorithme de Monte-Carlo. Chaque mutation apparaissant au sein de la population possède une certaine chance d'envahir la population : cette chance dépend de la différence d'adaptation entre les deux génotypes. La population est représentée par un seul génome et ce dernier est remplacé par le nouveau dès lors qu'une mutation parvient à fixation.

81. Daniel M Weinreich, Suzanne Sindi, and Richard A Watson. Finding the boundary between evolutionary basins of attraction, and implications for Wright's fitness landscape analogy. *J. Stat. Mech.*, 2013(01):P01001, January 2013

Taille du génome

Pour finir sur un cas pratique simple, la taille du génome est un critère important à plusieurs niveaux et pouvant réserver des surprises intéressante [82].

Très prosaïquement tout d'abord, la taille du génome limite la capacité d'un organisme à se reproduire puisque la vitesse de réplication de l'ADN – et le nombre de ribosomes – est limitée. Toutes choses égales par ailleurs, un organisme dont le génome est court pourra donc se reproduire plus vite et dominera la population. C'est sans doute pour cette raison que les bactéries disposent d'un génome très condensé avec très peu de parties non utilisées. Ceci n'est cependant vrai que dans si la réplication de l'ADN est un facteur limitant, ce qui est sans doute le cas chez les bactéries mais probablement pas chez les organismes pluri-cellulaires – qui possèdent d'ailleurs un génome majoritairement non codant [83].

Même dans le cas où la vitesse de réplication n'est pas limitante, la taille du génome a une influence importante sur le plan d'évolution. Ainsi, une chaîne courte est plus aisée à répliquer en ce sens qu'il y a moins de sites potentiels d'erreur. Il est par exemple démontré qu'un organisme ayant un taux d'erreur ponctuelle μ par site et par génération ne peut survivre s'il doit conserver plus de μ^{-1} sites. Au dela de ce taux connu comme le seuil d'erreur [84], la probabilité que chaque enfant ait au moins une erreur devient trop grande et la population finit par accumuler les mutations délétères ; la sélection n'étant pas suffisante pour maintenir la population fixée sur le meilleur génotype [85].

Ceci implique donc que contrairement à une idée reçue, la sélection ne choisira pas nécessairement le meilleur génotype qu'elle ait formée, mais plutôt la région de l'espace des génotypes $\mathcal{G}$ où la population aura la meilleure fécondité moyenne. Cet effet est connu comme la 'survie des plus plats' [86]. Il y a donc un intérêt certain à choisir le taux de mutation μ le plus faible possible.

82. Bérénice Batut, David P Parsons, Stephan Fischer, Guillaume Beslon, and Carole Knibbe. In silico experimental evolution: a tool to test evolutionary scenarios. *BMC Bioinformatics*, 14(Suppl 15):S11, October 2013

83. E. V. Koonin. *The logic of chance, the nature and origin of biological evolution*. FT Press Science, 2011

84. Christof K Biebricher and Manfred Eigen. The error threshold. *Virus Research*, 107(2):117–127, February 2005

85. La population choisira alors le meilleur génotype ne nécessitant la conservation que d'un nombre adéquat de sites.

86. Claus O Wilke, Jia Lan Wang, Charles Ofria, Richard E Lenski, and Christoph Adami. Evolution of digital organisms at high mutation rates leads to survival of the flattest. *Nature*, 412(6844):331–333, 2001

Mais d'un autre côté, un taux de mutation important permet d'explorer plus efficacement $\mathcal{G}$ ce qui, notamment en cas de variation brutale de l'environnement, peut se révéler salutaire [87]. On sent d'ores et déjà une interconnection forte entre le génome et le plan d'évolution. On voudrait que ce dernier puisse à la fois trouver de nouvelles solutions et conserver voire améliorer celles déjà existantes, ce qui demande bien une adéquation entre le génome et les mutations de ce dernier.

87. Ivana Bjedov, Olivier Tenaillon, Benedicte Gerard, Valeria Souza, Erick Denamur, Miroslav Radman, François Taddei, and Ivan Matic. Stress-induced mutagenesis in bacteria. *Science*, 300(5624):1404–1409, 2003

Notons pour finir que pour trouver la meilleure solution d'un problème donné, les algorithmes génétiques ne sont généralement pas la meilleure des possibilités [88]. En effet, tout comme l'évolution, ces derniers cherchent à maximiser le gain cumulé sur l'ensemble de la trajectoire ce qui peut conduire à des maxima locaux desquels il est difficile de s'extraire [89]. De plus, il manque à ce domaine une théorie explicative permettant par exemple de pouvoir donner pour un problème donné la structure du génotype et le plan d'évolution le plus adapté. Pour autant, ces algorithmes permettent de trouver des solutions relativement efficaces en des temps raisonnables et sont donc fortement utilisés lorsque ce type de solutions convient.

88. Melanie Mitchell, John H Holland, and Stephanie Forrest. When will a genetic algorithm outperform hill climbing? In J D Cowan, G Tesauro, and J Alspector, editors, *NIPS*, pages 51–58. Advances in Neural Information Processing Systems, 1994

89. J.H. Holland. *Adaptation in natural and artificial systems*. The MIT Press, 1992

Conclusion

L'évolution naturelle est un phénomène inévitable dont les grandes lignes sont aisées à discerner mais dont les détails sont particulièrement complexes. Elle regroupe différents mécanismes : sélection, neutralité, variation ; dont les importances respectives sont difficiles à identifier et à quantifier mais chacun laisse sa trace dans les organismes biologiques que nous observons. Si l'on veut utiliser l'évolution pour expliquer une propriété telle que l'architecture d'une protéine, il ne faut donc pas se demander quelle est son but ou son utilité mais quels mécanismes peuvent l'avoir produit et dans quelles conditions.

Pour cela, les algorithmes génétiques sont évidemment des outils de choix, surtout lorsque l'on peut les combiner avec des modèles simples d'objets biologiques.

Modèles d'évolution de protéines

> Let's say you're walking around and you find a watch on the ground. As you examine it, you marvel at the intricately complex interweaving of its parts, a means to an end. Surely you wouldn't think this marvel would have come about by itself. The watch must have a maker. Just as the watch has such complex means to an end, so does nature to a much greater extent. Just look at the complexity of the human eye. Thus we must conclude that nature has a maker too !
>
> William Paley, *Natural Theology*, 1802

Comprendre quelles sont les propriétés particulières qui distinguent les protéines naturelles des simples séquences aléatoires de nucléotides est une étape importante dans la compréhension des protéines en général et de leur évolution en particulier. Pour ce faire, il est intéressant de proposer des modèles simples permettant de se forger une intuition et de tester les propriétés génériques que l'on peut attendre d'une chaîne d'hétéropolymères sélectionnée pour des propriétés particulières telles qu'un repliement rapide, une conformation donnée ou une fonction particulière.

On l'a signalé précédemment, les protéines naturelles ne constituent qu'une proportion extrêmement réduite de l'ensemble des séquences possibles. Mieux encore, parmi toutes les conformations accessibles à une séquence d'acides aminés, les protéines naturelles ne semblent en utiliser qu'un nombre très restreint [90]. Lors de leur évolution, les protéines semblent donc s'être essentiellement concentrées autour d'un ensemble de séquences et de conformations très particulier.

90. C Chothia. Proteins. One thousand families for the molecular biologist. *Nature*, 1992

Quelles sont les raisons de ce choix particulier ? Est-il nécessaire ou contingent ? Pour répondre à ce type de questions, deux méthodes principales reposant toutes les deux sur l'utilisation de modèles simplifiés ont été explorées.

La première ne met pas directement en jeu l'évolution – même si cette dernière est toujours présente dans l'esprit de ces études. Elle cherche à déterminer les différences entre une chaîne aléatoire et une protéine structurée et à déterminer ainsi les propriétés qui sont valorisées par la sélection. Elle offre de plus un angle d'attaque pour comprendre la relation entre l'espace des séquences et celui des conformations.

La seconde méthode utilise explicitement l'évolution en mettant en œuvre une dynamique plus ou moins proche de celle des protéines naturelles pour étudier l'histoire évolutive d'une séquence.

On cherche ainsi à comprendre la réponse d'une population lorsqu'elle est soumise à une pression de sélection parfaitement contrôlée.

Ces deux méthodes utilisant majoritairement des modèles de protéines sur réseau, il nous semble donc nécéssaire d'en fournir une présentation rapide avant de détailler les différentes tentatives et les résultats de ces deux approches.

Modèles sur réseau

La première proposition d'utilisation d'un modèle sur réseau pour étudier le repliement des protéines remonte à 1967 dans un article de Kron *& al.* paru dans *Molekulyarnaya Biologiya* ; l'article est aujourd'hui pratiquement inaccessible. L'idée fut reprise par Taketomi, Ueda et Gō [91] en y ajoutant l'utilisation de simulations Monte-Carlo pour étudier la dynamique de ce repliement. L'exemple sera ensuite abondamment repris, notamment par Shakhnovich, en modifiant quelques éléments pour mettre en avant tel ou tel aspect des protéines.

91. H Taketomi, Y Ueda, and N Gō. Studies on protein folding, unfolding and fluctuations by computer simulation. *International journal of peptide and protein research*, 1975

L'idée maîtresse fut donc de fournir un modèle de protéine particulièrement adaptée à l'outil informatique puisqu'il permet une étude exhaustive des conformations de la chaîne. Pour cela on part d'une chaîne de longueur L composée d'acides aminés que l'on traitera comme des billes. Ces billes sont contraintes à demeurer sur les sites d'un réseau – généralement carré ou cubique – et la chaine suit les liens de ce réseau ainsi que le montre la figure (9). La protéine est donc vue comme un hétéro-polymère effectuant une marche auto-évitante sur les mailles d'un réseau simple.

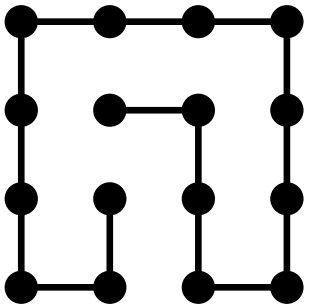

Figure 9: Exemple de modèle de protéine sur réseau en dimension 2 pour une chaîne de 16 sites.

L'énorme avantage est que l'on peut alors énumérer exhaustivement de telles marches. Par exemple, les marches auto-évitantes d'une chaîne $L = 20$ en 2 dimensions sont de l'ordre de 10^8 au plus [92]. Tandis qu'il n'existe que 802075 marches compactes de longueur 27 en dimension 3 (telle que celle de la figure 10) [93]. On peut donc suivre la dynamique de repliement tout en s'assurant par ailleurs que la conformation finale est bien celle d'énergie minimale. On peut même dresser le paysage énergétique complet de la protéine pour une séquence quelconque.

92. Estimation personnelle par simulation.

93. P D Williams, D D Pollock, and R A Goldstein. Evolution of functionality in lattice proteins. *Journal of Molecular Graphics and Modelling*, 19:150–156, 2001

Si l'on note σ_i et r_i respectivement la nature et la position de la bille i, le hamiltonien le plus général est de la forme :

$$\mathcal{H}(\sigma_i, r_i) = \sum_i E_0(\sigma_i) + E_1(\sigma_i, \sigma_{i+1}) + \sum_{ij} E_2(\sigma_i, \sigma_j, r_i, r_j), \qquad (14)$$

où le premier terme $(E_0 + E_1)$ décrit l'énergie de la séquence avant le repliement tandis que toutes les interactions dues à ce dernier sont décrites par le terme E_2. Deux types d'approximations permettent de simplifier l'étude de ce modèle.

Gō-like – Une approximation consiste à présumer l'existence d'une conformation de plus basse énergie, puis à construire les interactions de manières à favoriser cette dernière. Par exemple en af-

fectant une énergie très faible lorsque les différentes variables sont dans la conformation naturelle et une énergie plus ou moins aléatoire plus élevée lorsque ce n'est pas le cas. Cette méthode permet donc surtout d'étudier les propriétés d'une conformation donnée : sa stabilité, sa dynamique, etc. et d'évaluer ainsi la différence entre une séquence aléatoire et une séquence naturelle.

Interactions – L'autre approximation consiste à choisir une description plus ou moins décimée des interactions puis à étudier la dynamique d'une séquence particulière en suivant ces règles (par Monte-Carlo par exemple). Les études choisissent souvent d'ignorer l'énergie de la chaîne ($E_0 + E_1$) et de limiter E_2 aux interactions de courtes portées. On peut alors décrire les acides aminés selon deux grandes catégories, typiquement polaire ou hydrophobe (modèle HP) [94], auquel cas E_2 se limite à choisir la valeur des quelques paires possibles. Ou utiliser le jeu des vingts acides aminés naturels et choisir pour E_2 des énergies mesurées expérimentalement comme la matrice de Miyazawa-Jernigan [95]. On peut aussi choisir de défavoriser plus ou moins fortement la présence d'acides-aminés sur le même site si l'on s'intéresse à la dynamique du repliement comme dans l'approche de SHAKHNOVICH *et al.* [96].

Les verres de spins et les modèles de repliement de l'ARN ont aussi été utilisés pour ce type d'études comme nous le verrons dans la suite de ce chapitre, mais ils forment plutôt l'exception que la règle, aussi ne les présenterons nous pas *in extenso*.

94. K A Dill. Theory for the folding and stability of globular proteins. *Biochemistry*, 1985

95. S Miyazawa and R L Jernigan. Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. *Macromolecules*, 1985

96. Eugene Shakhnovich, G Farztdinov, A M Gutin, and Martin Karplus. Protein folding bottlenecks: A lattice Monte Carlo simulation. *Phys. Rev. Lett.*, 67(12):1665, 1991

Caractérisation des séquences et conformations naturelles

Les premiers travaux utilisant des modèles de protéines sur réseau pour chercher à déterminer s'il existe des conformations possédant des propriétés de repliement particulières datent des années du début des années 90. Par exemple, l'article de SHAKHNOVICH cité ci-dessus se propose d'étudier, à l'aide d'un modèle Gō-like couplé à une simulation de type Monte-Carlo, la dynamique du repliement d'une chaîne de 27 acides aminés sur un réseau 3D telle que celle de la figure (10) en se centrant sur des interactions à courte portée.

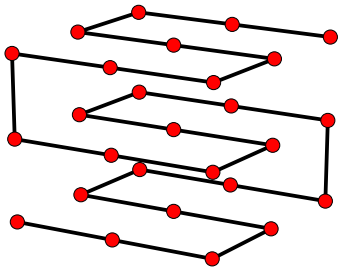

FIGURE 10: Exemple de modèle de protéine sur réseau, ici une structure cible compacte formée de trois couches indépendantes. En pratique cette structure est très difficile à atteindre car elle ne présente pas de chemin de repliement simple dans l'espace des configurations.

Cette étude conclut qu'il existe bien des conformations dont le repliement est facilité. La dynamique de repliement de ces dernières s'effectuent en deux étapes, d'abord une compaction rapide de la chaîne puis un équilibrage plus lent vers la structure native. Ceci n'est possible que grâce à une structure hiérarchisée qui n'est évidemment pas sans rappeler les structures secondaires et tertiaires des protéines et la séparation en domaines des protéines naturelles.

Pour faire le lien entre cette dynamique et la séquence des protéines, les modèles d'interactions sont cependant nécessaires. Ils présentent en effet l'avantage de ne pas introduire d'hypothèse sur la nature du paysage énergétique ni sur l'existence – ou l'unicité – d'une conformation fondamentale. Ceci permet donc d'effectuer

une approche « plus naturelle », notamment lors de l'évolution ; nous le verrons à la prochaine section. Le défaut étant évidemment qu'il faut alors parcourir exhaustivement l'ensemble des structures possibles pour retrouver la structure fondamentale, ce qui limite de telles études à des chaines courtes de l'ordre de la vingtaine ou de la trentaine d'acides aminés selon les cas.

Cette approche a par exemple permis de confirmer la présence de structures particulières correspondant à l'état de plus basse énergie d'un grand nombre de séquences [97]. Ce type de travaux a amené à construire la notion de *conformations concevables* (*designable* en anglais) pour désigner une structure facilement accessible. La concevabilité d'une conformation désigne le nombre de séquences différentes possédant cette dernière comme conformation de plus basse énergie.

Comme le nombre de configurations compactes est limitée, on peut déterminer rapidement la conformation d'énergie minimale correspondant à une séquence donnée. Cette approche permet d'investiguer les relations entre la séquence et la structure de manière détaillée, par exemple en construisant explicitement les réseaux de mutations neutres – ici celles qui ne modifient pas la conformation fondamentale, [98] ce qui permet de lever le voile sur les propriétés génériques de l'espace des génotypes. La structure de ce dernier peut alors être utilisée pour regarder l'influence du *plan de mutation* sur la répartition des séquences par exemple [99].

Mais l'on peut aussi utiliser cette approche pour retrouver des propriétés géométriques fondamentales des protéines. Les travaux de Trinquier et Sanejouand [100] utilisent un modèle HP en trois dimension en vue d'une meilleure caractérisation de ces structures concevables. Il s'agit ici avant tout d'expliquer la géométrie mais aussi la séquence de ces conformation et en particulier d'éclairer la relation entre les différentes séquences correspondant à une même structure. Ils montrent par exemple une préférence marquée pour les structures dont les résidus centraux sont hydrophobes. Une séparation qui n'est pas sans rappeler la séparation entre le cœur et la surface des protéines.

D'autres études sur des modèles similaires montrent que le nombre de boucles de faible taille au sein de la conformation est un bon indice de la facilité des chaînes aléatoires à posséder cette conformation comme fondamentale [101], permettant ainsi de caractériser la concevabilité des protéines naturelles.

Bien qu'elle ne soit pas toujours expressément utilisée, l'évolution est toujours au cœur de ces travaux. La concevabilité est en effet souvent considérée comme l'un des paramètres clé des protéines naturelles [102]. Au-delà de l'avantage évident d'une telle structure qui possède plus de séquences – et donc plus de chance d'être trouvée aléatoirement – la taille importante du domaine neutre dans l'espace des séquences fournit un nouvel avantage sélectif à ce type de séquences puisque l'on peut modifier profondément la séquence tout en gardant une structure stable [103] et donc en demeurant po-

97. Hao Li, Robert Helling, Chao Tang, and Ned Wingreen. Emergence of preferred structures in a simple model of protein folding. *Science*, pages 666–669, 1996

98. E Bornberg-Bauer and H S Chan. Modeling evolutionary landscapes: mutational stability, topology, and superfunnels in sequence space. In *PNAS*, 1999

99. Y Xia and M Levitt. Roles of mutation and recombination in the evolution of protein thermodynamics. In *PNAS*, 2002

100. G Trinquier and Y H Sanejouand. New proteinlike properties of cubic lattice models. *Phys. Rev. E*, 1999

101. Jeremy L England and Eugene I Shakhnovich. Structural Determinant of Protein Designability. *Phys. Rev. Lett.*, 90(21):218101, May 2003

102. R Helling, H Li, R Mélin, J Miller, and N Wingreen. The designability of protein structures. *Journal of Molecular Graphics and Modelling*, 2001

103. Konstantin B Zeldovich, Igor N Berezovsky, and Eugene I Shakhnovich. Physical Origins of Protein Superfamilies. *Journal of Molecular Biology*, 357(4):1335–1343, April 2006

tentiellement fonctionnelle. De plus, il est prouvé que des structures concevables sont généralement plus symétrique, se replient plus efficacement et sont plus stables [104].

104. R Mélin, H Li, N S Wingreen, and C Tang. Designability, thermodynamic stability, and dynamics in protein folding: a lattice model study. *The Journal of chemical physics*

Évolution directe de modèle de protéine

Les modèles de la section précédente nous ont donc permis de mettre en valeur les séquences et les structures possédant les propriétés minimales (stabilité, repliement rapide, etc.) leur permettant de former des protéines. Il est cependant nécéssaire de vérifier quelles sont les séquences qui émergent effectivement lors d'un processus de sélection et c'est là que les modèles d'évolution directe, souvent dérivés des algorithmes génétiques, entrent en action.

Ces modèles se sont tout d'abord principalement concentrés sur l'étude de propriétés proches du repliement. Les deux premiers modèles que nous présentons cherchent ainsi à faire évoluer les résidus d'une conformation donnée afin de rendre cette dernière native en maximisant la différence d'énergie entre la conformation fondamentale et le premier état dégénéré, une propriété censée garantir non seulement la stabilité mais aussi un repliement rapide de la protéine.

Dans le cas d'un modèle sur réseau de type HP [105] : on observe alors l'apparition d'une ségrégation entre les résidus polaires d'un côté et hydrophobes de l'autre. On retrouve à nouveau la séparation entre le cœur et la surface des protéines évoquée ci-dessus. L'immense avantage de ce modèle très simple est qu'il permet d'obtenir une approximation analytique du diagramme des phases indiquant la stabilité d'une structure donnée en fonction de la température et d'une « température de sélection » [106]. Cette « température » étant un paramètre fictif jouant un rôle similaire à la température et quantifiant les rôles relatifs de la sélection et des variations.

105. E I Shakhnovich and A M Gutin. Engineering of stable and fast-folding sequences of model proteins. *PNAS*, 90(15):7195–7199, 1993

106. Sharad Ramanathan and Eugene Shakhnovich. Statistical mechanics of proteins with "evolutionary selected"sequences. *Phys. Rev. E*, 50(2):1303, 1994

Des modèles de verre de spin ont également été proposés pour étudier la dynamique évolutive du repliement des protéines [107], en particulier la façon dont l'évolution cherche à façonner le paysage énergétique des configurations possibles de manière à faciliter le repliement de la protéine vers sa configuration fondamentale. C'est un genre de problèmes pour lequel les verres de spins avaient déjà été largement utilisés dans la physique de la transition vitreuse.

107. S Saito, M Sasai, and T Yomo. Evolution of the folding ability of proteins through functional selection. *PNAS*, 94(21):11324, 1997

En nous éloignant quelque peu des protéines, on trouve de nombreux modèles destinés à étudier les relations entre le phénotype, le génotype et l'évolution. Ces modèles servent à investiguer des notions telles que la neutralité, la robustesse et l'évoluabilité dont nous reparlerons plus en détail par la suite. On peut, par exemple, citer les travaux de Fontana et Schuster [108] utilisant un modèle de repliement de l'ARN. Ceci permet à partir d'un exemple relativement évolué de *plan de développement* et reposant sur une base physique, d'étudier la notion de neutralité, définie ici comme la taille typique de l'ensemble des génotypes donnant un même phénotype. Mais aussi de mieux comprendre la notion intuitive de

108. Walter Fontana and Peter Schuster. Shaping space: the possible and the attainable in RNA genotype-phenotype mapping. *Journal of Theoretical Biology*, 194(4):491–515, 1998

proximité entre deux phénotypes. En effet, deux structures complètement différentes peuvent en théorie être la conformation d'équilibre de deux séquences relativement similaire [109], et inversement deux séquences très différentes peuvent donner la même conformation. Ce type de travaux permet de quantifier l'importance de la dérive neutre, c'est-à-dire la dispersion d'une population dans la composante neutre de l'espace de génotypes, pour accélérer l'évolution de nouvelles fonctions.

Un autre exemple intéressant est le modèle de verre de spin de Sakata *& al.* [110] dans lequel les auteurs utilisent une double simulation de Monte-Carlo pour étudier l'influence des fluctuations dans l'évolution. L'idée est de choisir un modèle de verre de spin déterminé par les couplages J_{ij} ; de faire évoluer ces couplages avec une température T_J pour maximiser la corrélations entre un sous-ensemble prédéfini des spins ; les spins suivant eux même une dynamique d'équilibre beaucoup plus rapide danslaquelle J_{ij} est fixe mais les spins σ_i sont des variables dynamiques avec une température d'équilibre T_S. On a donc là aussi une relation complexe entre génotype et phénotype qui conduit à l'évolution d'un jeu de couplage permettant à basse température $T_S < T_S^{c2}$ de protéger les spins cibles du bruit thermique de la simulation et à plus basse température encore $T_S < T_S^{c1}$ d'éviter complètement la phase vitreuse du système en construisant un paysage énergétique dépourvu de frustration.

Enfin, un point mérite d'être ajouté pour mentionner l'utilisation de simulation toujours plus détaillées, de type tous-atomes bien que ces dernières se soient essentiellement concentrées sur la prédiction de la structure et le problème du repliement en délaissant l'étude de l'évolution. On peut cependant citer l'étude de l'évolution de la résistance au choc thermique par un modèle Gō-like détaillant tout les atomes des protéines étudiées [111] et l'évolution de structure de type ribosome par des simulations détaillées [112].

109. S Dalal, S Balasubramanian, and L Regan. Protein alchemy: changing β-sheet into α-helix. *Nature Structural & Molecular Biology*, 1997

110. Ayaka Sakata, Koji Hukushima, and Kunihiko Kaneko. Funnel Landscape and Mutational Robustness as a Result of Evolution under Thermal Noise. *Phys. Rev. Lett.*, 102(14):148101, 2009

111. I N Berezovsky and E I Shakhnovich. Physics and evolution of thermophilic adaptation. *PNAS*, 102(36):12742–12747, 2005

112. A Schug and W Wenzel. An Evolutionary Strategy for All-Atom Folding of the 60-Amino-Acid Bacterial Ribosomal Protein L20. *Biophysical Journal*, 90(12):4273–4280, June 2006

Fonction et Évolution, au delà du repliement

Tous ces travaux se sont essentiellement centrés sur la question du repliement. Au début des années 2000, de nouveaux modèles se sont intéressés à l'évolution de propriétés plus proches du caractère fonctionnel des protéines, généralement autour de la question de la fonction la plus simple : la liaison.

On peut par exemple sélectionner les séquences non pas pour une structure compacte mais pour des propriétés comme l'existence d'une simple poche de liaison – c'est-à-dire un site vide au cœur de la protéine – dans une chaîne 2D [113], un précurseur possible de la fonction de liaison. Ici, les auteurs cherchent surtout à étudier la connectivité du paysage évolutif – la façon dont les différentes chaînes fonctionnelles s'agencent entre elles – afin de décrire la géométrie de l'espace des génotypes.

Un autre modèle simple de liaison en 2D décrit la liaison comme

113. Benjamin P Blackburne and Jonathan D Hirst. Evolution of functional model proteins. *J. Chem. Phys.*, 115(4):1935, 2001

l'interaction d'une chaîne de 4 acides-aminés (la cible ou ligand) sur l'un des côté d'une protéine de longueur 16 préalablement repliée en un carré [114]. De façon intéressante, cette étude ne montre pas de différence qualitative dans la dynamique évolutive d'une population lorsque l'on sélectionne uniquement pour le repliement ou pour le repliement doublé d'une certaine fonction. Ce type de modèle peut aussi servir à étudier la relation entre les contraintes de stabilité et de fonction. Il est ainsi montré qu'il est plus facile de faire évoluer une fonction lorsque la contrainte sur la stabilité est progressivement relachée [115].

Une autre manière d'introduire la fonctionnalité – ici encore sous forme de liaison – est de demander la liaison ou la non-liaison des différentes protéines constituant un génome simplifié. On crée ainsi une dynamique plus riche et plus proche de l'évolution réelle de la protéine. Dans leur article, HEO *& al.* [116] utilisent ainsi trois protéines semblables à celle de la figure (10) avec des acides aminés sur chaque site en demandant que les deux premières forment un dimère tandis que la troisième doit rester isolée. On a donc des interactions positives et négatives entre les différents composants du génome permettant de construire un paysage évolutif dépourvu d'hypothèses *a priori* mais présentant des propriétés de frustration et d'interaction très intéressantes. On peut alors y tester l'influence du taux de mutation et questionner la notion d'*épistasie* [117] ou même d'espèce.

114. P D Williams, D D Pollock, and R A Goldstein. Evolution of functionality in lattice proteins. *Journal of Molecular Graphics and Modelling*, 19:150–156, 2001

115. Jesse D Bloom, Claus O Wilke, Frances H Arnold, and Christoph Adami. Stability and the evolvability of function in a model protein. *Biophysical Journal*, 86(5):2758–2764, 2004

116. M Heo, L Kang, and E I Shakhnovich. Emergence of species in evolutionary "simulated annealing". *PNAS*, 106(44):18638–18643, 2009

117. Le terme d'épistasie désigne les non linéarité du *plan de développement* : l'effet d'une mutation en un site donné dépendant *a priori* du reste du génome.

Ces modèles permettent donc d'explorer à l'aide de programmes rapides permettant de tester un grand jeu de paramètres différents, les propriétés de nombreuses séquences ou conformations différentes et de mettre en valeur les propriétés particulières de celles trouvées dans la nature. Ces simulations sont certes à très gros grains mais présentent l'avantage de porter sur des temps longs ou un grand nombre de séquences différentes, ce qui permet d'adopter des points de vue très généraux et d'aborder ainsi des propriétés très génériques qui sont souvent laissées de côté dans les études plus détaillées.

Les liens que ces méthodes permettent de tracer entre le paysage des génotypes et celui des phénotypes (c'est-à-dire le *plan de développement*) indiquent aussi des topologies très particulières qui vont avoir tendance à favoriser un certain type de séquence et de structure. En effet l'existence de quelques structures regroupant un grand nombre de séquences différentes apparaît de manière robuste à travers les différents modèles, que ce soit avec ou sans la présence de fonctions.

On peut donc commencer à répondre aux questions posées au début de ce chapitre. Les structures et séquences naturelles possèdent bien des propriétés singulières que l'on peut lier à un avantage évolutif. La dynamique de l'évolution semble de plus fournir naturellement certaines propriétés génériques telle que la séparation des

acides aminés entre un cœur hydrophobe et une surface hydrophile.

Pourtant en se concentrant sur le repliement, ces études laissent de côté un certains nombre de questions essentielles. Tout d'abord pourquoi les protéines se replient-elles, puisqu'il n'y a jamais – dans la nature – de sélection pour le repliement ? Que se passe t-il lorsqu'une protéine doit réaliser plusieurs fonctions simultanément ? L'étude des interactions entre les différentes protéines et de leur influence sur l'évolution telle qu'esquissé par Heo me parait aussi être un champ d'étude essentiel aussi bien pour la conception de protéines que pour la compréhension des forces fondamentales à l'oeuvre dans le vivant. On peut aussi espérer que les prochains développements de la puissance de calcul disponible rapprocheront encore les simulations de la réalité [118] et permettront l'étude de fonction plus complexes que la simple liaison à un ligand unique.

Pourtant, il s'avère que des fonctions pouvant être modélisées simplement comme l'allostérie manquent encore cruellement d'une interprétation évolutive. C'est ce que nous nous proposons de faire dans la seconde partie de ce tapuscrit. Nous devons cependant avant cela expliquer la notion de secteur et discuter de l'information que l'on peut d'ores et déjà extraire à partir de l'analyse détaillée des séquences génétiques.

118. Valentina Tozzini. Coarse-grained models for proteins. *Current Opinion in Structural Biology*, 15(2):144–150, April 2005

Corrélations, secteurs, comment lire une séquence ?

> [These old biologists are] far too unconscious of the fact that the solutions to these problems are in the first place statistical, and in the second place statistical, and only in the third place biological.
>
> Lettre de Pearson à Galton au sujet de la composition du Comité pour l'Évolution de la Royal Society, le 12 février 1897.

Les récents développements des méthodes de séquençage permettent de déterminer toujours plus rapidement le profil génétique de nombreuses espèces. Ces progrès ont permis d'identifier de nombreuses protéines présentes au sein d'une même espèce ou d'espèces différentes et partageant une grande similarité de séquence, signe qu'elles partagent toutes une origine évolutive commune. Les protéines de ces familles réalisent toutes des fonctions proches et permettent donc de comparer différentes solutions trouvées par l'évolution à un ensemble de problèmes fortement similaires.

Pour certaines de ces familles, on connait la séquence de plusieurs milliers de représentants. On peut dès lors chercher à aligner ces séquences à la façon de la figure (11), ce qui permet de faire facilement ressortir les similarités et les différences entre les séquences. Cet outil, appelé alignement de séquences multiples (MSA pour *Multi Sequence Alignment* en anglais), permet lorsqu'il est conjugué à des méthodes développées pour la physique statistique, de déduire de nombreuses informations sans jamais avoir à considérer la physique réelle de la famille de protéine considérée.

Ceci présente donc un grand avantage puisque les détails des interactions physiques entre les différents acides aminés sont souvent difficiles à déterminer et demandent par la suite une grande puissance de calcul pour pouvoir être utilisées.

La principale difficulté est alors de séparer ce qui relève de la phylogénie de ce qui relève de la fonction. Les similarités peuvent en effet provenir d'une origine commune sans avoir aucune conséquence sur la fonction de l'objet : mon frère et moi avons les yeux bruns parce que nous sommes de proches parents par exemple ; ou relever d'une réelle importance au regard de l'évolution : nous avons tous les deux des yeux parce que c'est utile !

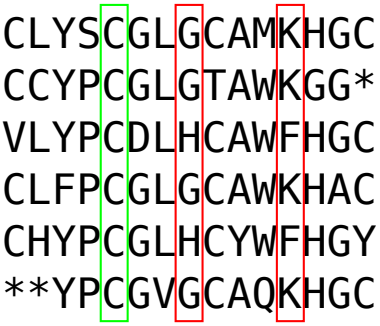

Figure 11: Exemple figuratif d'alignement de séquences, le site encadré en vert semble particulièrement conservé tandis que les sites encadrés en rouge semblent corrélés.

Alignement de séquences : quelle information ?

Supposons que l'on dispose d'un grand nombre N de séquences correspondant à une même famille de protéines. On notera S^i les différentes séquences et S^i_α l'acide aminé en position α dans cette séquence.

Pour chaque site α, on peut alors avoir une estimation de la fréquence d'apparition de l'acide aminé A :

$$f(S_\alpha = A) = \frac{1}{N}\sum_i \delta(S^i_\alpha, A), \tag{15}$$

où δ est la fonction qui vaut 1 si ses deux arguments sont égaux et 0 sinon.

Ceci conduit naturellement à une estimation de l'entropie par site [119] :

$$E_\alpha = -\sum_A f(S_\alpha = A)\log(f(S_\alpha = A)). \tag{16}$$

Intuitivement, l'entropie d'un site caractérise l'importance de ce dernier pour la protéine. Il peut être essentiel pour le repliement, pour la fonction, etc. Un site dont l'entropie est très basse, c'est-à-dire un site qui présente toujours le même acide aminé au sein de toutes les séquences étudiées laisse en effet à penser que ce dernier remplit un rôle essentiel pour l'une de ses fonctions [120]. Comme le taux de mutation est indépendant de la position, la présence répétée d'un même acide en un même site peut être due à la sélection de ce dernier : c'est-à-dire l'élimination de toutes les séquences qui ne le comporte pas. Pour autant, cette méthode ne suffit pas à décrire l'architecture d'une protéine.

Le meilleur moyen de s'en apercevoir et de construire une protéine *de novo* en se basant sur les probabilités p_α estimées précédemment. En générant puis en synthétisant des séquences aléatoires possédant les mêmes probabilités par site que les séquences naturelles, on s'aperçoit que ces dernières sont très loin de reproduire la fonction des séquences natives. En fait, la plupart de ces séquences artificielles s'avèrent même incapable de se replier correctement [121].

Nous avons en effet implicitement supposé dans notre démarche que ces sites étaient indépendants les uns des autres. Que la présence de tel ou tel acide aminé à une position n'influait en rien sur les probabilités du reste de la chaîne. Or une fois celle-ci repliée, les sites entrent en contact et il est nécéssaire pour cela que les appariements soient réalisés de façon adéquate. C'est là un problème de taille car nous devons estimer les probabilités dans un espace beaucoup plus grand. Il n'y a que 20 acides aminés mais il y a 400 paires différentes. Il faut donc avoir recours à des ensembles de séquences beaucoup plus importants, idéalement : $N \gg 400$. Heureusement ce chiffre est désormais atteint pour de nombreuses familles de protéines.

Le premier outil auquel on peut penser est alors l'information mutuelle. L'information mutuelle entre deux sites α et β, définie-

119. P S Shenkin and B Erman. Information-theoretical entropy as a measure of sequence variability. *Proteins: Structure, Function, and Genetics*, 1991

120. Du moins, si la fonction est effectivement identique pour toutes les séquences. Si la fonction est différente, il peut être intéressant de regarder l'information mutuelle entre un site et la fonction !

121. Michael Socolich, Steve W Lockless, William P Russ, Heather Lee, Kevin H Gardner, and Rama Ranganathan. Evolutionary information for specifying a protein fold. *Nature*, 437(7058):512–518, September 2005

par :

$$I_{\alpha\beta} = \sum_{AB} f(S_\alpha = A, S_\beta = B) \log\left(\frac{f(S_\alpha = A, S_\beta = B)}{f(S_\alpha = A) f(S_\beta = B)}\right), \qquad (17)$$

quantifie l'information obtenue sur un site par la connaissance du second.

Plus $I_{\alpha\beta}$ est grand, plus les deux sites sont corrélés. Dans notre cas, cela indique qu'il existe une forte corrélation entre la nature des acide aminés entre les deux sites. Pour autant cela ne signifie pas qu'il y ait un contact entre les deux sites ni que cette corrélation puisse permettre de tirer une quelconque information sur les caractéristiques physiques de la protéine. En effet, cette corrélation peut n'être qu'un effet de sous-échantillonnage, montrer que α et β sans être en contact partagent de nombreux voisins ou même indiquer une corrélation fonctionnelle à longue portée au sein de la protéine.

Comment déchiffrer une séquence

Deux méthodes permettent cependant de remonter à des informations pertinentes.

La première consiste à déterminer l'ensemble de couplages évolutifs [122] le plus probable étant donné les corrélations mesurées effectivement. Cette méthode dite de l'*analyse de couplage direct* permet ainsi de retrouver les couplages entre acides aminés et pourrait considérablement faciliter le problème du repliement chez les protéines naturelles.

L'autre technique consiste à déterminer les groupes de sites, c'est-à-dire les sites ayant une forte tendance à être corrélés au sein du groupe et peu avec l'extérieur. Ceci permet d'identifier les secteurs : les groupes d'acides-aminés partageant des caractéristiques évolutives communes.

L'analyse directe des couplages (DCA)

L'analyse directe des couplages est un exemple de problème inverse qui fut proposé par WEIGT *et al* [123] afin de supprimer les corrélations indirectes dans l'information mutuelle. Cette méthode part de l'hypothèse que l'essentiel des corrélations entre sites provient des contacts physiques et cherche ainsi à retrouver une partie de ces derniers par une analyse de séquence… sans reposer sur la physique réelle de la protéine étudiée. Les auteurs font remarquer que l'observation d'une corrélation donnée entre deux sites provient d'un ensemble de connexion plus ou moins courtes et plus ou moins nombreuses dont le couplage direct entre les deux sites n'est qu'un élément ainsi qu'il est schématiquement représenté sur la figure (12).

Heureusement, ce type d'effet peut être simulé et donc corrigé. L'idée générale étant de faire une hypothèse sur les couplages, d'en simuler l'effet sur les corrélations puis de corriger progressivement

122. Notons ici que le lien entre les couplages physiques ou contacts et les couplages évolutifs n'est pas encore clairement compris.

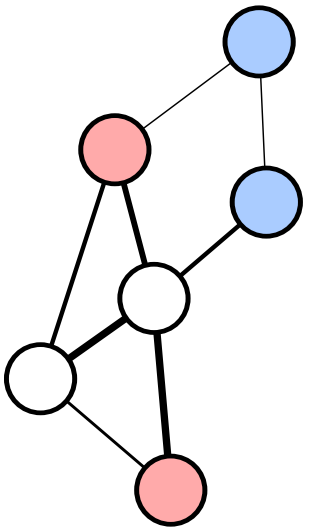

FIGURE 12: Si l'épaisseur des traits dans ce graphe indique la force des interactions entre sites, il est possible que les deux sites rouges, bien que sans contact direct, soient plus corrélés que les deux sites bleus.

123. M Weigt, R A White, H Szurmant, J A Hoch, and T Hwa. Identification of direct residue contacts in protein–protein interaction by message passing. *PNAS*, 106(1):67–72, 2009

l'hypothèse initiale pour obtenir le meilleur ajustement des données et de la simulation.

Dans la pratique, cette méthode repose sur de nombreuses techniques provenant de la physique statistique. Une méthode d'entropie maximale montre qu'un modèle de Potts est le modèle minimal suffisant pour analyser les couplages au sein des protéine [124]. On part donc du Hamiltonien :

$$\mathcal{H}(S) = -\sum_{\alpha\beta} J_{\alpha\beta} S_\alpha S_\beta - \sum_\alpha h_\alpha S_\alpha, \tag{18}$$

dont il s'agit maintenant de déterminer les différents paramètres $J_{\alpha\beta}$ et h_α.

À partir de (18) il est possible de déterminer les statistiques de l'ensemble produit par ce modèle : $f(S_\alpha = A)$ et $f(S_\alpha = A, S_\beta = B)$ puis d'ajuster les paramètres afin de retrouver les valeurs expérimentales. L'analyse des couplages ainsi déterminés, $J_{\alpha\beta}$ permet de déterminer les positions les plus fortement contraintes deux à deux, celles qui ont le plus de chance d'être en contact si notre hypothèse est correcte.

Il reste un problème de taille, déterminer les statistiques des séquences pour un jeu de couplages donné est une étape coûteuse en terme de temps de calcul et qui nécessite donc un choix astucieux quand à la méthode utilisée. Dans leur article, Weigt *et al* proposent d'utiliser un algorithme à passage de message qui permet de réduire drastiquement la dépendance dans la taille de la séquence analysée. Il sera montré par la suite que l'on peut même utiliser une approche de champ moyen ou d'autres techniques pour recouvrir cette information [125].

La méthode montre alors de bons résultats et permet de détecter des contacts directs, retrouvant ainsi les paires de contact déterminées expérimentalement. Détail intéressant, cette méthode permet également de retrouver les contacts fonctionnels entre deux protéines – ce qui est le cas dans le premier article cité – et pas seulement au sein d'une même protéine.

L'analyse statistique des couplages (SCA)

Plutôt que d'utiliser les statistiques précédemment déterminées pour regarder les contacts entre les différentes positions,une autre méthode propose de regarder les grands groupes d'acides aminés du point de vue de l'évolution.

L'objectif est alors de chercher à minimiser l'influence de la phylogénie pour faire apparaître les corrélations à longue portée au sein de la protéine. Ceci fait ressortir des ensembles particuliers appelés *secteurs évolutifs* dont on cherchera ensuite à déterminer les propriétés d'un point de vue structurel et fonctionnel.

Une première utilisation d'une telle technique a permis à Lockless et Ranganathan de déterminer un nombre important de caractéristiques du domaine PDZ [126], révélant notamment l'existence d'un ensemble d'acides aminés fortement couplés au site

124. William Bialek and Rama Ranganathan. Rediscovering the power of pairwise interactions. *arXiv*, q-bio.QM, December 2007

125. Faruck Morcos, Andrea Pagnani, Bryan Lunt, Arianna Bertolino, Debora S Marks, Chris Sander, Riccardo Zecchina, José N Onuchic, Terence Hwa, and Martin Weigt. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *PNAS*, 2011

126. Le domaine PDZ désigne une famille de domaines de protéine très variée – 260 exemplaires rien que chez l'homme – d'une centaine d'acides aminés de long et présente chez les bactéries, les eukaryotes et les virus. Il est principalement impliquée dans la régulation grâce à ses propriétés de liaison et d'allostérie.

actif [127]. Ce sont ces ensembles qui recevront plus tard le nom de *secteurs*.

La méthode utilisée est inspirée de l'analyse des marchés financiers [128]. Dans le cas des marchés financiers, l'idée est d'utiliser les corrélations entre la baisse et la hausse de différentes actions pour retrouver les secteurs d'activités économiques : l'ensemble des banques, des constructeurs automobiles ou d'armement ayant tendance à montrer davantage de corrélations avec les actifs du même secteur qu'entre secteurs.

Dans le cas des protéines [129], cette méthode révèle un ou plusieurs groupes d'acides aminés, connectés entre eux dans la structure tertiaire mais largement éclatés le long de la structure primaire. Ces acides aminés coévoluent fortement au sein d'un groupe mais très faiblement entre deux groupes. Les acides aminés ont d'ailleurs tendance à ne jamais appartenir à plus d'un groupe. Autrement dit, il n'y a pas de recouvrement entre les différents secteurs d'une même protéine. Lorsqu'on les soumet à une analyse expérimentale, ces groupes sont de plus généralement associés à des caractéristiques différentes de la protéine : stabilité, liaison, etc.

La technique de SCA proprement dite repose sur l'étude des modes dominants de la matrice des corrélations C. Pour cela, on définit dans un premier temps :

$$C_{\alpha\beta}^{AB} = f(S_\alpha = A, S_\beta = B) - f(S_\alpha = A) f(S_\beta = B) \tag{19}$$

la corrélation entre les deux sites pour une paire d'acides aminés donnée.

Afin de faire ressortir les éléments les plus pertinents, on introduit la pondération

$$P_\alpha^A = \log\left(\frac{f(S_\alpha = A)}{f(A)} \frac{1 - f(A)}{1 - f(S_\alpha = A)}\right), \tag{20}$$

où $f(A)$ est la fréquence d'apparition de A au sein de l'ensemble du génome. Ceci permet d'accorder une grande importance aux cas où la distribution d'un acide aminé sur un site donné est très différente de celle du reste du génome. C'est une façon de répondre à la question de la séparation entre phylogénie et fonction que nous avons soulevée au début de ce chapitre. Une fois ce biais introduit, il ne reste qu'à en calculer la moyenne quadratique, ce afin d'éviter la compensation des termes positifs et négatifs [130] :

$$C_{\alpha\beta} = \sqrt{\sum_{AB} \left(P_\alpha^A P_\beta^B C_{\alpha\beta}^{AB}\right)^2}. \tag{21}$$

Les différents modes de cette matrice contiennent, entremêlées de manière complexe, des informations sur la phylogénie des espèces dont les séquences sont étudiées et les corrélations évolutives des différents acides aminés entre eux. Démêler ces deux informations est encore aujourd'hui une question ouverte [131]. Certains éléments sont cependant relativement bien compris.

127. Steve W Lockless and Rama Ranganathan. Evolutionarily conserved pathways of energetic connectivity in protein families. *Science*, 286(5438):295–299, 1999

128. Vasiliki Plerou, Parameswaran Gopikrishnan, Bernd Rosenow, Luís A Nunes Amaral, Thomas Guhr, and H Eugene Stanley. Random matrix approach to cross correlations in financial data. *Phys. Rev. E*, 65(6):066126, June 2002

129. Najeeb Halabi, Olivier Rivoire, Stanislas Leibler, and Rama Ranganathan. Protein Sectors: Evolutionary Units of Three-Dimensional Structure. *Cell*, 138(4):774–786, 2009

130. O Rivoire. Elements of coevolution in biological sequences. *Phys. Rev. Lett.*, 2013

131. Olivier Rivoire, Kimberly A Reynolds, and Rama Ranganathan. The Structure of Evolutionary Constraints in Proteins. *En Préparation*

La théorie des matrices aléatoires permet, par exemple, de déterminer précisément à partir de quel mode les informations ne sont plus distinguable du bruit de fond en utilisant la théorie des distribution des valeurs propres d'une matrice aléatoire. Mais dans notre cas et notamment à cause des effets de taille finie, il est sans doute plus pertinent de générer un ensemble de séquences conservant les mêmes propriétés statistiques mais dépourvues de corrélations et d'en calculer les valeurs propres afin d'obtenir un modèle nul. Cet ensemble aléatoirisé de nos données permet ainsi de déterminer le seuil au-delà duquel une mesure est significative.

Dans le cas de la famille de la protéase à sérine S1A étudié par Halabi *et al* [132], l'étude des trois modes ainsi identifiés révèle qu'ils sont directement liés à des propriétés fonctionnelles de la protéine tel que le repliement, l'énergie de liaison ou encore l'activité catalytique. Ces secteurs présentent de plus une grande indépendance à la fois statistique (non recouvrement des modes) et fonctionnelle (des mutations dans un secteur n'affectent que la propriété lié à ce secteur).

L'une des caractéristiques des secteurs est de former, dans la structure tertiaire, un ensemble connecté qui semble transporter un signal ou une information. Le rôle des secteurs est donc particulièrement intéressant pour explorer l'allostérie qui correspond mieux qu'aucune autre fonction au transport d'information d'un site actif à un autre. Il est en particulier montré que dans le cas de la dihydrofolate réductase, les acides aminés de surface les plus proches du secteur responsable de la fonction forment des points privilégiés pour l'apparition d'un contrôle allostérique de cette dernière [133].

132. Najeeb Halabi, Olivier Rivoire, Stanislas Leibler, and Rama Ranganathan. Protein Sectors: Evolutionary Units of Three-Dimensional Structure. *Cell*, 138(4):774–786, 2009

133. Kimberly A Reynolds, Richard N McLaughlin, and Rama Ranganathan. Hot Spots for Allosteric Regulation on Protein Surfaces. *Cell*, 147(7):1564–1575, 2011

À la suite de ces deux exemples, on pourrait vouloir continuer l'analyse et s'intéresser aux corrélations entre trois sites, quatre sites et ainsi de suite. Deux objections sont cependant à considérer. D'une part, plus on désire augmenter la portée des corrélations, plus il est nécéssaire de posséder un très large ensemble de séquence afin d'éviter de travailler dans un régime dominé par le bruit d'échantillonnage. Or si ce régime est atteint pour les corrélations à un point et peut être contourné pour les corrélations à deux points en se limitant aux familles possédant de nombreux représentants au sein de chaque génôme, on est encore loin de pouvoir constituer des alignements de taille suffisante pour passer à l'étape suivante.

De plus une étude expérimentale [134] montre que des séquences synthétiques reproduisant uniquement les statistiques sur un et deux sites d'une famille de protéine se replient dans la structure de cette dernière dans 28% des cas alors que ce taux est proche de zéro si l'on ne prend que l'information sur un site [135]. Ceci permettant d'avancer que les corrélations à deux points fournissent une grande partie de l'information qu'il est possible de récupérer par analyse statistique.

134. Michael Socolich, Steve W Lockless, William P Russ, Heather Lee, Kevin H Gardner, and Rama Ranganathan. Evolutionary information for specifying a protein fold. *Nature*, 437(7058):512–518, September 2005

135. Dans cette expérience, les séquences naturelles de la famille ne se replient elle même correctement que dans 67% des cas.

Discussion

Dans les deux méthodes d'analyse des couplages présentées ci-dessus, rien n'indique *a priori* un lien entre la statistique des séquences et la fonction biologique. Les protéines répondent en effet à différentes contraintes qu'il convient de mettre en perspective :

Séquence : une protéine est nécessairement composée d'acides aminés appartenant à un alphabet restreint, ce qui permet d'effectuer un traitement statistique de cette information,

Évolution : puisque c'est un tel processus qui permet de construire les protéines fonctionnelles et de les conserver,

Structurel : c'est-à-dire ayant trait au repliement spatial et temporel depuis la séquence primaire jusqu'à la structure quaternaire,

Fonction : le rôle joué par la protéine au sein de l'organisme.

Chacune de ces contraintes a des conséquences différentes et demande une analyse et des outils particuliers, mais ils sont tous évidemment interconnectés. Ainsi, une protéine ne peut réaliser sa fonction qu'une fois correctement repliée et si une séquence ne remplit pas sa fonction – par exemple par suite d'une mutation – elle ne tardera pas à disparaître de la population sous le coup de l'évolution. Chacune de ses perspectives nous apporte donc certes des indices et un éclairage sur les autres, mais il faut garder à l'esprit qu'il n'y a jamais d'équivalence stricte. Une forte information mutuelle peut ainsi être le résultat d'une contrainte fonctionnelle, structurelle ou le simple produit d'une évolution aléatoire.

Pour comprendre les résultats de ces deux méthodes, il convient cependant de bien dissocier les différentes notions de couplages. Les *couplages physiques* désignent les interactions existants entre les acides aminés, elles sont *a priori* de courtes portées et sont les responsables du repliement et de la fonction. Les *couplages évolutifs* désignent l'ensemble des contraintes pesant sur la nature de deux acides aminés au cours de l'évolution. Ce sont ces derniers qui provoque l'apparition de corrélation entre les sites lors du processus d'évolution.

La méthode SCA cherche à déterminer les groupes de sites couplés évolutivement tandis que la DCA suppose que les couplages évolutifs sont le produit des couplages physiques et cherche à déterminer ces derniers. Cependant, tout couplage physique n'est pas un couplage évolutif. En particulier, les acides aminés en contact mais appartenant à deux secteurs différents semblent avoir évoluer pour être indépendants. De tels contacts ne peuvent donc probablement pas être retrouvés par la méthode DCA.

Développer des modèles permettant de mieux comprendre les liens entre les couplages physiques et les couplages évolutifs semble désormais essentiel pour mieux manipuler ces outils. En particulier, déterminer si un couplage évolutif est nécessairement un couplage physique et quelles sont les conditions pour que cette implication soit vraie ou fausse serait un résultat intéressant dans le

cadre de l'étude statistique des séquences.

Nos questions sur l'architecture des protéines sont maintenants établies sur de bonnes bases. Nous avons montré qu'une fraction des acides aminés seulement d'une protéine apporte une contribution majeure pour les fonctions de cette dernière. Ces derniers peuvent être mis en valeur par l'évolution, directement par mutation ou plus indirectement des techniques physiques comme les réseaux élastiques, dans chaque cas nous nous référerons à ce sous ensemble sous la notion de secteur(s). De plus, ces secteurs possèdent des architectures et des propriétés de hiérarchie intéréssante. Nous allons maintenant nous intéresser à l'évolution pour tenter de mieux comprendre l'apparition de ces secteurs au sein des protéines.

Nous utiliserons pour cela des modèles simplifiés de protéines, dans la lignée de ceux présentés dans cette première partie. Nous mettrons toutefois de côté le problème du repliement pour nous concentrer sur la notion de fonction.

Deuxième partie

Travaux personnels

Protéines parcimonieuses et modulaires

> Complex objects are produced by evolutionary processes in which two factors are paramount : the constraints that at every level control the systems involved, and the historical circumstances that control the actual interactions between the systems.
>
> F. Jacob, *Evolution and Tinkering*, Science, 1977.

Nous avons vu dans les chapitres précédents que l'évolution naturelle avait produit chez les organismes biologiques des structures hautement hiérarchiques. Pourtant, la plupart des algorithmes génétiques peinent à produire une telle modularité [136]. Il semble notamment difficile pour un algorithme d'apprendre seul à discerner les différents éléments qui composent un problème complexe.

De nombreuses hypothèses ont donc été proposées afin d'expliquer l'émergence de modularité au sein du vivant et permettre ainsi de construire des algorithmes d'apprentissages plus efficaces. Deux grands courants peuvent être distinguées.

D'une part certains chercheurs attribuent cette modularité à la structure du *plan d'évolution* [137], par exemple grâce aux possibilités ouvertes par les recombinaisons génétiques : ce seraient les transferts de gènes horizontaux chez les bactéries ou les enjambements (*cross-over* en anglais) chez les organismes sexués qui seraient à l'origine de la structure modulaire observée. Si les gènes peuvent être transportés par blocs d'un organisme à un autre, il y existera en effet une sélection pour les gènes (ou les groupes de gènes suffisamment proches le long du génome) capables d'opérer de manière autonomes qui formeront ainsi les modules observés.

Les autres invoquent la modularité de l'environnement lui-même comme une explication de l'apparition de la hiérarchie dans le vivant [138]. Le terme d'*environnement modulaire* désigne ici la structure du problème auquel fait face le système, on parlera de modularité quand ce dernier peut être séparé en différents sous-problèmes plus simples. L'idée étant que la structure du problème conditionne en quelque sorte la structure de la solution.

Pour prendre un exemple simple, si le sucre et la lumière étaient toujours présents simultanément, un organisme évoluerait probablement un seul système pour répondre à ces deux évènements. Mais s'ils sont indépendants l'un de l'autre, leurs détections et leurs traitements correspondraient à deux fonctions séparées.

136. Nadav Raichman, Ronen Segev, and Eshel Ben-Jacob. Evolvable hardware: genetic search in a physical realm. *Physica A: Statistical Mechanics and its Applications*, 326(1-2):265–285, August 2003

137. R Calabretta, S Nolfi, D Parisi, and G P Wagner. Duplication of modules facilitates the evolution of functional specialization. *Artificial life*, 2000

138. Nadav Kashtan and Uri Alon. Spontaneous evolution of modularity and networks motifs. *PNAS*, 102:13773–13778, September 2005; and Evan A Variano and Hod Lipson. Networks, Dynamics, and Modularity. *Phys. Rev. Lett.*, 92(18):188701, May 2004

Ces deux réponses ne sont cependant pas antagonistes. En effet, le plan d'évolution des organismes vivants est lui-même un élément de ces organismes et ainsi soumis à sélection. Il parait donc vraisemblable que l'apparition de plans d'évolution modulaires soit elle-même une réponse à un environnement modulaire.

On peut cependant vouloir examiner l'une de ses deux propositions de façon isolée, car les mécanismes à l'œuvre dans l'évolution dépendent de l'échelle à laquelle on se place. Ainsi, si la structure modulaire des réseaux de gènes ou la formation de domaines peut être due à la façon dont les gènes se déplacent le long du ou des chromosomes, il n'est pas envisageable d'invoquer une telle explication pour l'apparition de secteurs dans un même domaine. En effet, comme le montre la figure (13), les acides aminés d'un secteur sont dispersés tout au long de la séquence du domaine, bien qu'il forme une structure continue dans la structure tertiaire. De plus, lorsque plusieurs secteurs sont présents au sein d'une même protéine, ces derniers sont enchevêtrés.

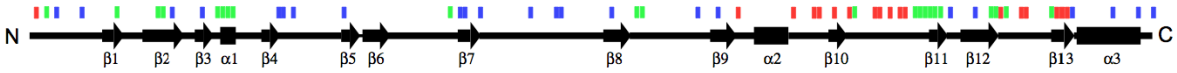

Figure 13: Représentation des 3 secteurs (en rouge vert et bleu) le long de la structure de la trypsine. Le caractère imbriqué et dispersé des secteurs est ici manifeste. (Reproduit de Halabi & al., *Cell*, 2009.)

Il apparait donc peu probable que la modularité des secteurs puisse s'expliquer par un mécanisme de mutation particulier. Ces derniers forment donc un objet de choix pour étudier l'influence de l'environnement dans l'apparition de modularité.

Avant d'aborder la question de la modularité, nous travaillerons longuement la question de l'apparition de *parcimonie* [139] sous l'évolution, c'est-à-dire de la concentration de la fonction dans une région restreinte de la protéine et l'apparition conjointe d'une zone neutre au sein de laquelle les mutations n'auront pas ou peu d'effet. Cette caractéristique nous parait en effet être un pré-requis important dans le développement de la modularité.

Notre hypothèse est donc que le caractère parcimonieux des protéines [140], manifesté par l'apparition des secteurs, est le résultat de l'évolution de ces dernières au sein d'un environnement particulier. Par environnement, nous désignons la suite de pressions de sélection imposées sur cette protéine en particulier. La question que nous nous posons est donc la suivante : « Étant donné un modèle de protéine, peut-on contrôler l'apparition d'une forme de parcimonie au sein de ce dernier en jouant sur l'environnement extérieur ? »

Nous cherchons donc un modèle inspiré des protéines pour lequel l'environnement peut forcer l'apparition de parcimonie. En particulier, notre modèle devrait montrer un caractère dépensier dans un environnement particulier et parcimonieux dans un autre.

Comme la modélisation de fonction de protéines réelles est largement en dehors des possibilités de calculs actuelles (sans parler des nôtres), nous proposons d'utiliser un modèle à gros grains

139. Nous utilisons le terme de "parcimonie" pour l'anglais *sparsity*, c'est-à-dire la propriété d'un système dont un grand nombre de paramètre sont nuls (ou proches de zéros).

140. Nous raisonnerons toujours dans la suite sur une protéine composée d'un seul domaine, les termes domaines et protéines sont alors équivalents.

permettant de réaliser des simulations portant sur des tailles raisonnables de population – de l'ordre de 500 à 1000 individus minimum – durant des temps relativement long, c'est-à-dire grands devant l'échelle de variation de l'environnement.

Nous avons choisi d'étudier un modèle d'allostérie pour plusieurs raisons. D'une part, c'est une fonction couramment rencontrée dans les protéines et pour laquelle la notion d'environnement variable est relativement aisée à définir puisqu'il existe deux ligands susceptibles de varier avec le temps. De plus, c'est une fonction qui n'a pas *a priori* de concentration géométrique puisque la corrélation entre les sites actifs peut mettre en jeu l'ensemble de la protéine. En effet, contrairement à une activité de liaison qui implique un rôle particulier d'une zone précise de la protéine : le site actif, l'allostérie de par sa nature de passer un message peut impliquer n'importe quel acide aminé avec une égale probabilité. Afin d'éliminer complètement l'influence de la geométrie, nous utiliserons une geométrie simplifiée plutôt que d'utiliser les cartes de contacts de protéines réelles, en nous plaçant ainsi dans le cas d'une symétrie maximale : toute brisure de symétrie sera nécessairement la trace d'un choix parcimonieux dû à l'environnement.

Nous avons évoqué précédemment l'utilisation de modèles gaussiens élastiques pour calculer l'activité allostérique de protéines réelles en soulignant combien ce modèle présente l'avantage d'être très facile à implémenter et peu coûteux en temps de calcul : c'est pourquoi nous choisirons d'utiliser ce type de modèle dans la suite de notre exposé. Notez cependant que le choix du modèle est peu important, l'utilisation de modèles de verre de spins décrit en appendice (p. 117) – ayant eux aussi étaient utilisés pour représenter des protéines – mène à des conclusions similaires. C'est d'ailleurs le choix de présentation que nous avions fait dans notre article [141].

141. Mathieu Hemery and Olivier Rivoire. Evolution of sparsity and modularity in a model of protein allostery. *Phys. Rev. E*, 91(4):042704, April 2015

Il apparait alors que sous des hypothèses très larges, un environnement variant dans le temps produira des protéines d'autant plus parcimonieuses que le temps caractéristique τ de ces variations est faible.

La cause de cette parcimonie est à chercher dans la capacité du système à répondre au changement d'environnement en produisant et en sélectionnant des mutations adaptées. Comme ces mutations sont aléatoires et non dirigées – c'est-à-dire qu'elles ne dépendent pas de l'environnement – le paramètre clé est la probabilité qu'un site donné mute au moins une fois parmi toute la population au cours d'une période de l'environnement.

Ainsi, une population de taille N dont le taux de mutation par site est μ possède un taux de mutation effectif par couplage de μN. C'est-à-dire que la probabilité d'observer une mutation en un site donné au cours d'une génération sur l'ensemble de la population est μN. La probabilité d'une mutation au cours d'une période est donc d'ordre $\tau\mu N$. Si cette quantité est petite devant 1, le modèle

n'est pas capable de suivre l'environnement et se retrouve perdu. Si cette quantité est très grande devant 1, le modèle s'adapte beaucoup plus vite que l'environnement ne change et on retrouve le cas d'un environnement constant. Dans le cas où $\tau\mu N \simeq 1$ le modèle se stabilise en n'utilisant pour la fonction qu'un sous ensemble de la protéine : il est ainsi capable d'adapter cette dernière en un temps raisonnable alors qu'il ne lui serait pas possible de s'adapter si tous les sites étaient mobilisés.

Il existe dans ce cas un compromis à trouver entre la capacité à s'adapter lors d'un changement d'environnement qui augmente d'autant plus que la région impliquée dans la fonction est réduite et la capacité à remplir la fonction la plus efficace possible entre deux changements qui demande au contraire une région la plus large possible.

Dans notre cas, ce paramètre joue principalement sur la largeur de la région fonctionnelle. Mais ce type de compromis peut se retrouver sur la longueur de cette dernière avec des effets de taille inverse comme décrit par notre modèle à une colonne décrit dans le chapitre suivant (p. 67).

Notons que ce type de compromis favorise fortement la capacité à s'adapter malgré le faible temps durant lequel cette dernière joue un rôle prédominant. En effet, l'étude temporelle de la réponse montre que si la population s'adapte rapidement après un changement d'environnement ; ce processus implique l'élimination des individus ayant tentés de développer une région plus large. C'est ce phénomène qui en bloquant l'évolution d'une région plus large provoque l'apparition de la parcimonie.

Enfin, pour en venir à la modularité, cette dernière apparaît dans notre modèle lorsque l'on demande au système de remplir plusieurs fonctions ou que la fonction est séparée en plusieurs taches distinctes. Cette condition n'est cependant pas toujours suffisante car comme nous l'avons signalé, les algorithmes génétiques ont tendances à produire des solutions non modulaire. Nous avons cependant montré que le détail de la statistique de l'environnement permet ou non à l'évolution de distinguer les sous fonctions facilitant ainsi l'évolution de solutions modulaires.

Nous nous attacherons particulièrement à montrer que ces effets reposent sur des hypothèses relativement souples même si les détails précis des résultats peuvent dépendre du modèle utilisé.

Nous commencerons par présenter une notion simplifiée de la parcimonie dans le cadre d'un modèle très simple, le modèle à une colonne, permettant de mettre en évidence les différentes forces en jeu et les questions qu'elles soulèvent ? Puis nous présenterons en détail le modèle gaussien et l'algorithme génétique que nous utiliserons pour développer quantitativement ces hypothèses. Nous prendrons aussi du temps pour parcourir les différentes variantes de ces modèles que nous avons testés afin de montrer que la plu-

part de nos choix ne sont pas essentiels et de souligner la généralité de cette explication. Nous discuterons ensuite de l'apparition de la modularité avant de montrer en quoi ce modèle permet de répondre aux différentes questions que nous nous sommes posées en abordant la notion de secteur. Nous terminerons en présentant les limites et les développements possibles d'un tel modèle.

Modèle à une colonne

> There are two principal ways to formulate mathematical assertions : Russian and French. The Russian way is to choose the most simple and specific case (so that nobody could simplify the formulation preserving the main point). The French way is to generalize the statement as far as nobody could generalize it further.
>
> Vladimir Arnold, *Arnold's Problems*, 2005

Pour vérifier, dans un premier temps, si l'évolution est capable de déterminer un compromis simple entre fonction et adaptation, prenons le modèle le plus simple qu'il nous soit possible d'imaginer tout en conservant une inspiration physique.

Nous cherchons à corréler ou anti-corréler deux points $A = \sigma_0$ et $B = \sigma_\ell$ en utilisant une chaîne de spins σ_i dont les couplages sont de force constante et de signe variable : $J = \pm 1$ comme représenté sur la figure (14). Pour une température non nulle $T \propto \frac{1}{\beta}$, il existe une certaine probabilité que le message soit transmis de manière incorrecte, le taux d'erreur ainsi introduit ne dépendra cependant que du nombre ℓ de couplages entre A et B.

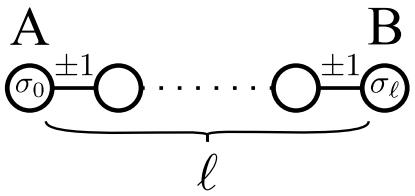

Figure 14: Schema de notre modèle simplifié d'allostérie. La fonction est de corréler ou d'anti-corréler les spins présents aux deux extrémités A et B de la chaîne.

En utilisant un modèle d'Ising 1D dont le hamiltonien est donné par :

$$H(\sigma) = \sum_{i=1}^{\ell} J_i \sigma_{i-1} \sigma_i, \tag{22}$$

on montre facilement que la corrélation entre les deux extrémités de la chaîne est donnée par :

$$< \sigma_0 \sigma_\ell >= \prod_{i=1}^{\ell} \tanh(J_i \beta). \tag{23}$$

Ce que nous choisirons comme base pour calculer la fécondité. Comme tout les couplages sont de même intensités, on a donc :

$$< \sigma_0 \sigma_\ell >= \pm \tanh(\beta)^\ell, \tag{24}$$

selon la parité du nombre d'interactions négatives.

Nous souhaiterions construire un modèle d'évolution sans toutefois recourir à un algorithme génétique. Chaque individu dispose donc directement d'un taux de croissance et la taille de population sera maintenue constante par un changement d'échelle à chaque pas de temps. Le taux de croissance est donné par :

$$\phi = 1 + a < \sigma_0 \sigma_\ell >= 1 \pm a \tanh(\beta)^\ell = 1 \pm s, \tag{25}$$

où a est une constante quantifiant l'importance de la fonction étudiée pour la sélection.

Pour simuler le changement d'environnement, nous demandons donc que les deux extrémités soient tantôt corrélées tantôt anti-corrélées, ce qui ne demande en fait qu'un changement de signe d'une des interactions le long de la chaîne, donc une seule mutation.

Il y a deux types de mutations : les mutations de longueur qui modifient le nombre de liens ℓ et les mutations sur chaque lien qui en modifient le signe. Ces taux de mutations respectifs seront notés μ_ℓ et μ_s. Si l'on se concentre tout d'abord sur les mutations de signe, la probabilité que le signe global de la chaîne change en une génération est donné par la probabilité d'un nombre impair de mutation. Si $\mu_s \ll 1$, on peut supposer que ce terme est dominé par le cas où il n'y a qu'une seule mutation et l'on a un taux effectif $\mu_{\text{eff}} = \ell \mu_s$.

On note $P(t)$ le vecteur décrivant la population à l'instant t. Pour l'instant on garde ℓ fixé et l'on ne s'intéresse qu'aux deux sous-populations constituées des individus adaptés p_a (dont la fécondité est $1+s$) et les non-adaptés p_{na} (dont la fécondité est $1-s$).

$$P(t+1) = \begin{pmatrix} p_a(t) \\ p_{na}(t) \end{pmatrix} \tag{26}$$

Chaque génération et représentée par la matrice d'évolution :

$$P(t+1) = M.P(t) = \begin{pmatrix} (1+s)(1-\mu_{\text{eff}}) & (1-s)\mu_{\text{eff}} \\ (1+s)\mu_{\text{eff}} & (1-s)(1-\mu_{\text{eff}}) \end{pmatrix} .P(t). \tag{27}$$

Un changement d'environnement inverse p_a et p_{na} et est donc décrit par la matrice

$$C = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}. \tag{28}$$

Le taux de croissance à long terme d'une telle population dans un environnement variant toutes les τ générations est simplement donné par la racine τ-ième de la plus haute valeur propre de la matrice $M^\tau C$ que nous noterons λ.

Deux questions peuvent se poser sur ce système :

— Existe t-il pour un τ donné, une valeur de ℓ qui maximise la croissance à long terme λ ?

— L'évolution converge t-elle naturellement vers cette valeur ?

Pour déterminer l'existence d'une valeur maximale $\ell^\star$, le plus simple est de calculer ces matrices numériquement pour une gamme de valeur raisonnable. Le calcul analytique n'est en effet ni élégant ni instructif et les deux paramètres cruciaux (τ et ℓ) prennent de toutes façons des valeurs discrètes.

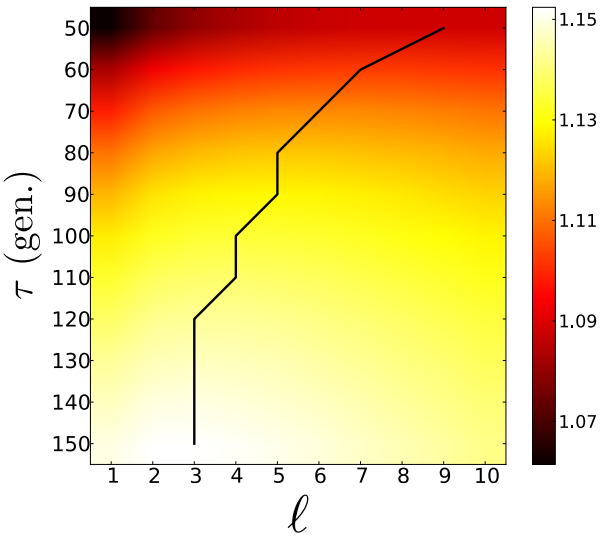

FIGURE 15: Carte de λ en fonction de ℓ et τ. Les paramètres choisis sont $\mu_s = 10^{-3}$, $a = 0.2$ et $\tanh(\beta) = .99$. La ligne noire indique la position du maximum pour τ fixé.

On obtient ainsi la figure (15) dans laquelle la ligne noire indique la valeur $\ell^\star$ pour laquelle λ est maximal à τ fixé. Il s'avère que cette valeur se comporte comme $\tau^{-3/2}$ ainsi que montré dans la figure (16).

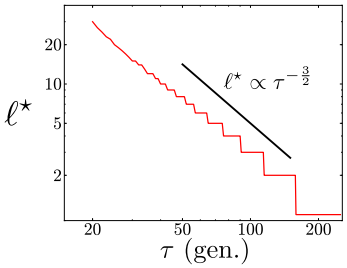

FIGURE 16: Comportement de $\ell^\star$ en fonction de τ pour des paramètres typiques, la loi de puissance est manifeste.

Pour des paramètres raisonnables, il existe donc une valeur de longueur permettant le meilleur compromis entre adaptation et efficacité. Remarquez que dans ce cas précis, la longueur ℓ diminue la fécondité mais augmente l'évoluabilité tandis que dans notre modèle principal, la largeur du canal augmentait la fécondité en diminuant l'évoluabilité. Ceci explique que notre paramètre optimal se comporte comme $\tau^{-3/2}$ tandis qu'il se comportait plutôt comme τ dans notre modèle de protéine.

On voit donc qu'un système évolutif en environnement variable peut jouer sur sa géométrie : tanten longueur qu'en largeur, pour s'adapter à un environnement complexe.

Il nous reste maintenant à vérifier dans quelle mesure une simulation de l'évolution aboutit naturellement sur cette valeur $\ell^\star$. Pour cela nous utilisons l'algorithme suivant :

— La population est décrite par un tableau donnant le nombre d'individus adaptés et non adaptés de taille ℓ dans la population pour ℓ allant de 0 à 20.

— Chaque groupe croit ou décroit proportionnellement à sa fécondité en utilisant une loi de Poisson, puis ces groupes sont soumis à mutation (là encore en suivant des lois de Poisson) avec les taux μ_{eff} et μ_ℓ

— Une étape de mise à l'échelle de la population élimine aléatoirement des individus au sein de la population afin de conserver une taille globale N_{pop} constante.

Cet algorithme présente l'avantage de permettre des simulations relativement rapides et de ne pas dépendre de la taille de population utilisées, ce qui permet de tester les cas des grandes tailles de population qui sont justement généralement difficiles à simuler.

De premier résultats présentés dans la figure (17) montrent que, si la convergence est presque parfaite lorsque la taille de population est importante et le taux de mutation μ_l est faible, c'est loin d'être le cas pour tous les choix de paramètres. Il existe en particulier des régimes dans lequel l'évolution est pratiquement bloquée sur la configuration initiale (ici $\ell = 5$) ou totalement perdue.

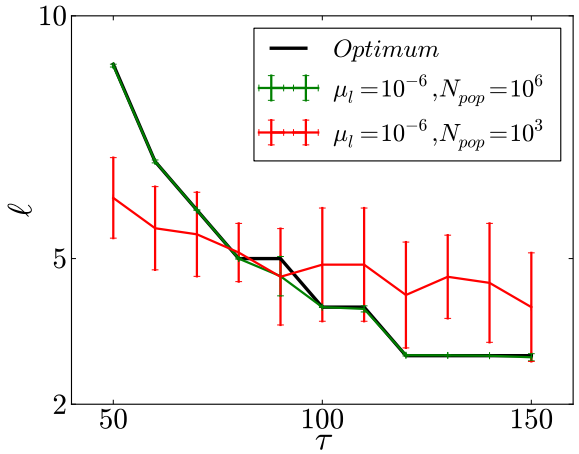

FIGURE 17: Exemple de résultats pour différents choix de paramètres. La ligne noire indique la valeur de l'optimum calculé précédemment.

Le meilleur moyen de vérifier quels jeux de paramètres mènent à une évolution efficace est de comparer l'écart à l'optimum pour différentes valeurs de μ_l et de N_{pop} comme illustré par la figure (18). On y voit clairement trois régions apparaître. Dans la première correspondant à $\mu_l N_{pop} \gg 1$ (en bleu en haut à droite), l'évolution est littéralement bloquée et on reste proche de la situation de départ $\ell = 5$. Dans l'immense majorité du reste des cas (en blanc), l'évolution est efficace et converge plus ou moins rapidement vers la solution optimum. Enfin, dans le coin supérieur gauche pour lequel N_{pop} est petit et μ_l important, l'évolution est à la fois dominée par une grande part de hasard mais on voit que les individus s'adaptant trop vite (ceux pour lesquels $\ell > \ell^\star$) dominent la population.

Ceci laisse à penser que les effets de taille finie pourrait même dans certains cas favoriser l'évoluabilité au dépend du taux de croissance à long terme !

Ce modèle simple permet de mettre en valeur les liens entre les propriétés évolutives et la géométrie d'un système et ouvre des pistes pour comprendre certains aspects complexe de l'évolution tels que la sélection de l'évoluabilité. En effet, les tailles de population pour lesquels des phénomènes particuliers apparaissent ne sont pas très éloignées des tailles effectives de population que l'ont

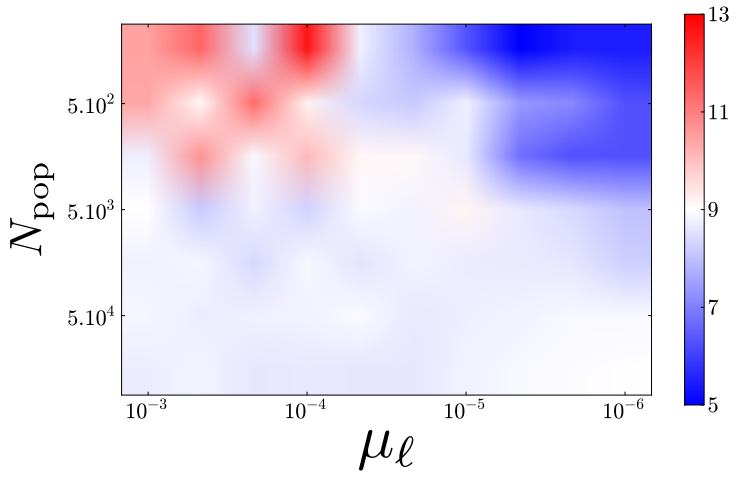

FIGURE 18: Carte de ℓ en fonction de μ_l et N_{pop}. Les paramètres choisis sont identiques aux précédent avec $\tau = 50$, pour cette valeur l'optimum est $\ell = 9$ en blanc sur la figure.

peut rencontrer chez certaines espèces [142].

142. M Lynch. The Origins of Eukaryotic Gene Structure. *Molecular Biology and Evolution*, 23(2):450–468, September 2005

Évolution d'un modèle d'allostérie

La perfection est atteinte, non pas lorsqu'il n'y a plus rien à ajouter, mais lorsqu'il n'y a plus rien à retirer.
Antoine de Saint-Exupéry

Afin de développer un modèle quantitatif pour l'étude de l'apparition de parcimonie et de modularité au sein des protéines, nous aurons besoin des éléments principaux d'un algorithme génétique.

Il nous faudra donc un génotype – c'est-à-dire une façon de décrire un individu. Ce dernier impliquera aussi le choix du plan d'évolution. Il nous faudra aussi un plan de développement, ici le moyen de passer du génotype à une évaluation de l'activité allostérique de la protéine, ce dernier devra s'inspirer au plus près des modèles d'allostérie existants tout en demeurant d'un temps de calcul modéré. Enfin nous devrons choisir une fonction de fécondité pour déterminer le nombre de successeurs d'un génotype donné dans la génération suivante.

Nous utiliserons l'hypothèse de non recouvrement des générations en posant $\sigma = 0$ dans l'équation (11).

Génotype et plan de développement

Nous nous proposons donc de travailler avec une géométrie simplifiée, celle d'un réseau carré enroulé autour d'un cylindre permettant de définir les contacts entre sites voisins. On utilisera généralement un réseau de 10×10 sites comme présenté sur la figure (19). Les sites sont notés s_i et les couplages entre deux sites voisins K_{ij}. Nous utiliserons ensuite un modèle de réseau élastique mais contrairement au cas classique, nous supposerons que la constante de raideur de chaque site peut varier. C'est l'ensemble des couplages $\{K_{ij}\}$ qui constitue alors notre génotype.

Nous ignorerons ici complètement le problème du repliement. Nous allons supposer que la structure tertiaire de la protéine est fixée et que les couplages que nous modifions ne font qu'en perturber la dynamique [143]. Notre hamiltonien comportera donc un terme de structure fixant la position de chaque site et des couplages en évolutions déterminant les modes libres et la dynamique du système. Comme s_i représente l'écart de l'acide aminé par rapport à sa position initiale, l'ajout d'un terme quadratique fixe liant chaque acide aminé à sa position est suffisant, la constante de couplage de

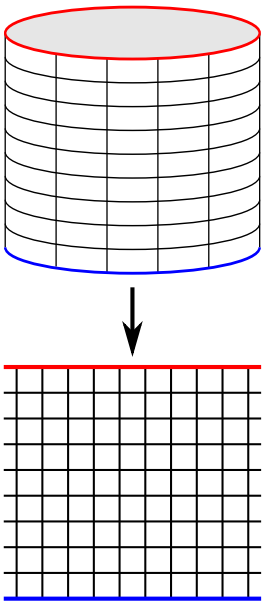

Figure 19: Géometrie de notre modèle de protéine en 3D et sa projection en 2D.

143. Cela revient en fait à supposer que notre évolution se déroule dans le réseau neutre correspondant à une structure hautement concevable donnée.

ce nouveau potentiel K_r sera supposé être identique pour tous les sites.

Afin de simuler les modifications de conformation de notre protéine sous l'effet de l'environnement, la position d'équilibre des sites peut être modifiée par un terme élastique de constante e_i analogue à un champ extérieur. Ce dernier indique l'influence d'un éventuel ligand ou du solvant par exemple.

Pour un génotype fixé, le hamiltonien de notre système est ainsi donné par :

$$\mathcal{H}(s_i|K_{ij}) = -\frac{1}{2}\sum_i K_r(s_i - e_i)^2 - \frac{1}{2}\sum_{ij} K_{ij}(s_i - s_j)^2, \tag{29}$$

où le premier terme indique la partie purement structurelle de la protéine et le second décrit la partie fonctionnelle qui sera notre objet d'étude.

En développant les carrés, on peut séparer les termes diagonaux et hors-diagonaux ce qui donne :

$$\mathcal{H}(s_i|K_{ij}) = -\sum_i \frac{1}{2}(K_r + \sum_j K_{ij})s_i^2 + K_r e_i s_i + \sum_{ij} K_{ij} s_i s_j. \tag{30}$$

Le terme en e_i^2 à été ignoré puisque, constant, il ne joue aucun rôle dans la dynamique du système.

Ce hamiltonien s'écrit beaucoup plus aisément sous une forme matricielle :

$$\mathcal{H}(s) = \frac{1}{2}s^T\tilde{K}s + K_r s^T e, \tag{31}$$

où $\tilde{K}$ est définie par $\tilde{K}_{ij} = -K_{ij}$ en dehors de la diagonale et $\tilde{K}_{ii} = (K_r + \sum_j K_{ij})$ sur la diagonale.

Pour un environnement e donné, il devient alors facile de déterminer l'énergie libre de la protéine. En effet, la fonction de partition :

$$Z(e,K) = \int \exp(-\beta\mathcal{H}(s|K,e))ds, \tag{32}$$

est réduit à une intégration gaussienne [144] :

$$\begin{aligned} Z(e,K) &= \int \exp(-\tfrac{1}{2}s^T\tilde{K}s - K_r s^T e)ds, \\ &= \sqrt{\tfrac{(2\pi)^n}{\det(\tilde{K})}}\exp\left(\tfrac{K_r^2}{2}e^T\tilde{K}^{-1}e\right), \end{aligned} \tag{33}$$

144. Dans le suite on pose $\beta = 1$ pour simplifier les calculs, ce terme pouvant toujours être absorbé dans notre définition des couplages K_{ij} et K_r.

sous réserve que cette expression soit bien définie. Dans ce cas l'énergie libre est donnée, à une constante près, par :

$$F(e,K) = -\ln(Z(e,K)) = \frac{1}{2}\ln\det(\tilde{K}) - \frac{K_r^2}{2}e^T\tilde{K}^{-1}e + \text{constante}. \tag{34}$$

Avant de poursuivre, nous aimerions vérifier quelles sont les conditions sur nos paramètres pour s'assurer que notre intégrale converge et fournit une valeur physiquement plausible, c'est-à-dire que $\tilde{K}$ est bien inversible et $\det(\tilde{K})$ soit positif.

Plaçons nous dans le cas où $|K_{ij}| < 1$, alors $\tilde{K}$ est la somme de deux termes $\overline{K}$ et $K_r\mathbb{I}_n$. Donc en notant λ une valeur propre de $\overline{K}$,

$\lambda + K_r$ est une valeur propre de $\tilde{K}$. De plus, la somme des valeurs absolues sur chaque ligne et chaque colonne de $\overline{K}$ est bornée par $2k$ où k est la connectivité maximale du réseau. Dans ces conditions, les valeurs propres sont elles aussi bornées par $2k$ et en choisissant $K_r > 2k$ on s'assure que $\tilde{K}$ est inversible et que son determinant est positif. Notre intégrale est ainsi assurée d'être définie pour tout jeu de couplage K_{ij}.

Dans la suite nous nous placerons toujours dans le cas où $|K_{ij}| < 1$ et $K_r = 10 > 2k$. Comme il est fait remarquer dans l'appendice consacré au verre de spin (p. 113), cette limite correspond à un cas de haute température et donc à une forte influence du bruit. Les chemins courts sont donc fortement favorisés pour porter la fonction, ce qui est l'une des limitations de ce modèle.

Le calcul de l'activité allostérique α est enfin réalisé en séparant le vecteur e en trois parties : le site actif, le site de régulation et le cœur. Comme le cœur sera toujours laissé libre (c'est-à-dire que l'on suppose que dans le cœur on a $e_i = 0$), on notera $e(\ell_1, \ell_2)$ la configuration avec le ligand ℓ_1 au site actif et ℓ_2 au site de régulation. Les ligands seront toujours pris uniformes et composés de 1 et de -1 en cas de présence d'un ligand et de 0 si le site est vacant. On a donc essentiellement deux ligands différents : le ligand positif et le ligand négatif que nous noterons $+1$ et -1 dans la suite.

L'énergie de liaison du ligand ℓ_1 au site actif est alors calculée par :

$$\begin{aligned} \Delta F(\ell_1, 0) &= F(e(\ell_1, 0), K) - F(e(0,0), K) \\ &= \tfrac{K_r^2}{2} \left(e(0,0)^T \tilde{K}^{-1} e(0,0) - e(\ell_1, 0)^T \tilde{K}^{-1} e(\ell_1, 0) \right), \end{aligned} \tag{35}$$

et l'activité allostérique par

$$\alpha(\ell_1, \ell_2) = \Delta F(\ell_1, \ell_2) - \Delta F(\ell_1, 0). \tag{36}$$

Notez que nous n'avons pas introduit le potentiel chimique des ligands en solution car de par la construction de l'activité allostérique, ces derniers se simplifient lors du calcul de α qui doit, par définition, être une propriété de la protéine indépendante de la solution.

Le principal coût en terme de temps de calcul lors du calcul de α vient donc de l'inversion de la matrice $\tilde{K}$. Mais cette matrice est creuse par construction et donc facilement inversible [145].

145. Dans le cas d'un réseau d'une centaine de sites, il faut compter de l'ordre de 10^{-2} seconde pour une évaluation de l'activité allostérique en utilisant les modules ordinaires de python.

Plan d'évolution

La seule contrainte que nous ayons pour l'instant posée sur notre génotype est $|K_{ij}| < 1$. On en vient donc assez naturellement à initialiser notre génotype en choisissant uniformément nos différents couplages dans l'ensemble $[-1, 1]$.

Le plan d'évolution doit reposer sur des hypothèses minimales. Idéalement, on voudrait pouvoir comparer différents choix de plan d'évolution afin de certifier que l'effet de concentration n'est pas

(ou du moins peu) dépendant de ce dernier. Nous souhaitons de plus exclure tout phénomène de recombinaison de large échelle, du type enjambement, pour nous concentrer sur des mutations locales afin de certifier que l'apparition de modularité est bien une propriété de l'environnement et non de notre plan d'évolution.

Biologiquement, c'est l'acide aminé d'un site qui est modifié par les mutations et on pourrait donc vouloir qu'une mutation affecte l'ensemble des couplages liés à un nœud. Bien qu'il puisse sembler plus fidèle, un tel choix est en fait restrictif. L'élément essentiel étant *une séparation entre les variables évolutives et les variables physiques*. On peut donc en fait utiliser n'importe quel type de mutation du moment que ce dernier est suffisamment local. Le cas extrême serait la modification d'un seul couplage à la fois par exemple.

Plusieurs arguments plaident en faveur d'un tel choix. D'une part la simplicité, modifier un seul couplage étant le choix le plus élémentaire. D'autre part la généralité, car le choix de la géométrie de la zone affectée ne doit pas affecter nos résultats. Enfin, un argument du type "renormalisation" pourrait montrer que toute modification locale est équivalente à la modification d'un unique couplage dans un système plus restreint *via* un changement d'échelle adapté. Comme nous cherchons à la fois la simplicité et la généralité, nous nous concentrerons donc sur des mutations couplage par couplage mais consacrerons un bref aparté au cas de mutation sur les sites.

Comment une mutation peut-elle affecter notre couplage ? De nombreux types de mutation sont envisageables :

Sans mémoire : Chaque mutation ré-initialise complètement le couplage en tirant une nouvelle valeur dans $[-1,1]$. $K' = \mathcal{U}([-1,1])$

Processus additif : Les mutations sont représentées par l'ajout d'une variable aléatoire au couplage existant. $K' = K + \mathcal{N}(0,\sigma_a)$

Processus multiplicatif : Les mutations sont représentées par la multiplication du couplage existant par une variable aléatoire. $K' = K \times \mathcal{N}(1,\sigma_m)$

Sans mémoire discret : En choisissant un ensemble de valeurs entre -1 et 1 puis en tirant tous nos couplages aléatoirement dans cet ensemble.

Des valeurs discrètes présentent de plus un grand intérêt théorique car elles permettent de construire des espaces de mutations munis d'une géométrie de notre choix afin de tester des théories particulières. Par exemple, le diagramme de la figure (20) montre un exemple sur lequel on peut distinguer des nœuds mutables (au centre du diagramme) et des nœuds fixes (sur les branches), on peut alors regarder quels types de nœuds sont préférés selon la nature de l'environnement.

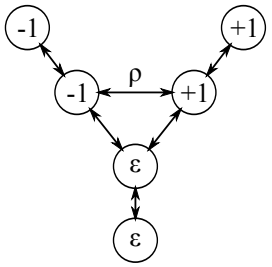

FIGURE 20: Exemple de graphe possible pour les couplages, on s'attend à voir un environnement constant utiliser davantage les noeuds terminaux (plus stables) qu'un environnement variable qui se concentrerait sur les trois noeuds fortement connectés et plus variables. $\epsilon << 1$ correspond à un couplage faible et ρ peut prendre n'importe quelle valeur (éventuellement différente pour chaque flèche).

Fécondité

Le choix de la fonction de fécondité est grandement facilité par le fait que notre phénotype se réduit à une seule quantité, qui plus est un nombre réel. Pourtant, nous sommes tout de même handicapé par le fait que ce nombre peut prendre une grande plage de valeur, tant positives que négatives. Afin d'éviter des effets de distorsion possibles avec une méthode de sigma-scaling, on préférera donc utiliser une simple stratégie élite plus rapide et plus robuste.

Ce choix correspond de plus à celui couramment fait dans la littérature [146]. Et il n'influe pas les résultats puisque le choix du *sigma-scaling* a été choisi pour l'article et mène à des conclusions similaires.

146. Nadav Kashtan and Uri Alon. Spontaneous evolution of modularity and networks motifs. *PNAS*, 102:13773–13778, September 2005

Notre algorithme est donc fin près, tel que représenté dans la figure (21).

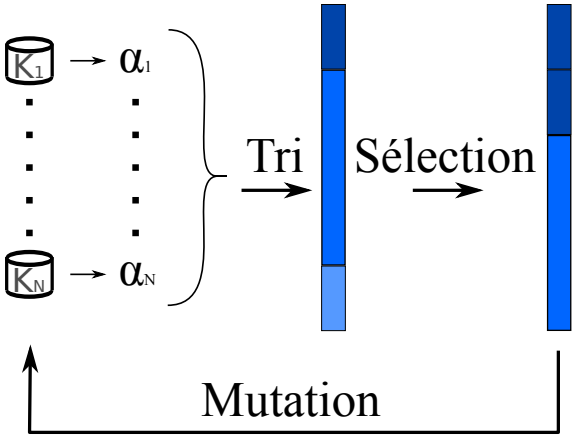

FIGURE 21: Schéma d'une génération de notre algorithme génétique (stratégie "élite").

Analyse d'un individu

Nous allons maintenant nous intéresser aux outils que nous mettrons en place pour étudier le comportement de notre système. Nous supposerons que nous disposons d'une population décrite par une liste d'individus, chacun étant caractérisé par la liste de ses couplages K_{ij}.

Le moyen le plus évident pour représenter un individu serait de représenter la force de ses couplages sur un réseau carré (le signe de chaque interaction n'est pas pertinent à représenter en raison des invariants comme ceux de la figure 22). Ceci permet d'obtenir un premier aperçu du résultat de l'évolution. On voit d'ailleurs déjà apparaître une forme de concentration en comparant le réseau de couplage pour le meilleur individu de la population après une évolution en environnement constant et en environnement variable comme le montre la figure (23).

Plus l'environnement varie rapidement (donc plus τ est petit), moins les couplages sont uniformément les plus important pos-

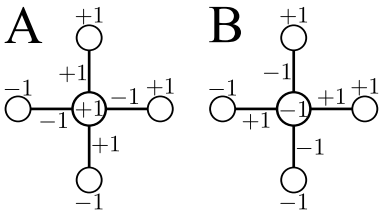

FIGURE 22: Exemple d'invariance rendant l'analyse des signes délicates dans notre modèle, les configurations **A** et **B** sont indiscernables du point de vue de l'énergie malgré les différences entre les signes des couplages.

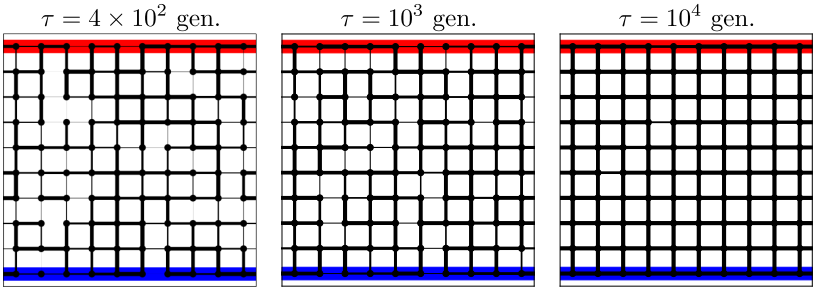

FIGURE 23: Couplages du génotype le plus adapté de la population pour différentes valeurs de période τ, l'épaisseur du trait est proportionnelle à l'intensité de chaque couplage (en valeur absolue). L'utilisation de couplages d'autant plus forts que la période est grande est manifeste, mais que peut-on dire de plus ?

sibles. On voit apparaître des groupes de couplages forts dans une partie du système tandis que le reste est difficilement identifiable. En effet, il n'existe que peu de sélection négative dans notre modèle, les couplages en dehors de la zone utilisée ne sont donc pas contraints à être les plus faibles possibles mais simplement dépourvus de sélection et prennent ainsi des valeurs aléatoires. Discerner cet aléatoire est difficile et nous aimerions disposer d'un outil plus précis pour mettre en exergue la concentration de la fonction dans telle ou telle zone.

En particulier, nous voudrions pouvoir identifier les couplages les plus importants pour la fonction. Pour ce faire, inspirons-nous des expériences réalisée en biologie ! Pour identifier les résidus les plus importants, une technique consiste à muter un par un les différentes résidus en analysant à chaque fois l'influence de chaque mutation sur la fonction. Certains sites apparaissent alors comme particulièrement sensibles pour l'efficacité de la protéine tandis que d'autres ne semblent jouer pratiquement aucun rôle.

En nous inspirant de cette méthode on définit le coût d'un couplage comme la différence (relative) d'efficacité allostérique induite par la suppression de ce couplage :

$$C(K_{ij}) = 1 - \frac{\alpha(K|K_{ij} = 0)}{\alpha(K)}, \tag{37}$$

où $K|K_{ij} = 0$ désigne l'ensemble K avec le couplage entre i et j "supprimé", c'est-à-dire réduit à zéro.

On obtient alors la figure (24) dans laquelle la concentration des couplages sur une zone géométrique particulière du réseau est particulièrement manifeste. Cette représentation est intéressante car elle permet de mettre en valeur l'aspect géométrique de la concentration. Les couplages essentiels ne sont pas dispersés aléatoirement au sein de la protéine mais regroupés en un canal continu qui n'est pas sans rappeler les secteurs au sein des protéines réelles.

Nous aimerions cependant disposer d'une grandeur simple permettant de rendre compte de la parcimonie d'une protéine ; plusieurs choix sont possibles. On pourrait tout d'abord compter le nombre de couplages « essentiels » c'est-à-dire dont le coût est supérieur à un seuil donné. Bien que cette analyse dépende de la valeur seuil choisie, cette dépendance est faible pour des choix raisonnables. Une valeur $C_s \simeq 5\%$ fournit par exemple d'excellents

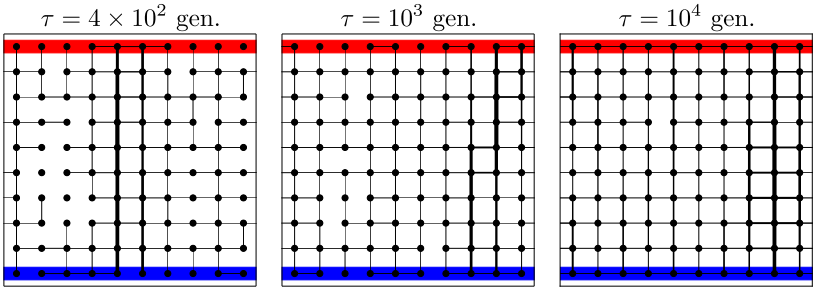

FIGURE 24: Coût des couplages du génotype le plus adapté de la population pour différentes valeurs de période τ, l'épaisseur du trait est proportionnelle au coût de chaque couplage. La concentration sur un bandeau est maintenant manifeste, la largeur de ce dernier semblant croître avec la période τ.

résultats comme le montre la figure (25). Une deuxième façon de déterminer la concentration d'une protéine est de s'intéresser à la variance de l'ensemble des couplages. Plus la protéine est concentrée, plus sa variance est forte car il y aura de fortes disparités dans les coûts des différents couplages (figure non montrée).

Il nous semble important de disposer d'une comparaison adéquate pour déterminer si une population est adaptée ou non. Nous choisirons pour cela comme modèle nul le meilleur individu d'une population dont les couplages sont tirés aléatoirement dans l'intervalle $[-1, 1]$. Au vu de notre modèle, ce choix correspond aussi au cas d'une population ayant été soumise à une sélection pour un seul environnement et sans introduction de variations – c'est-à-dire avec le plan d'évolution trivial.

La figure (26) montre que pour des valeurs de période τ très faible, la population n'est pas capable de s'adapter et son activité allostérique n'est pas supérieure à celle de notre modèle nul. Ceci indique la limite inférieure qu'il est possible d'atteindre – toutes conditions égales par ailleurs – pour laisser à la population le temps de développer une réponse mesurable.

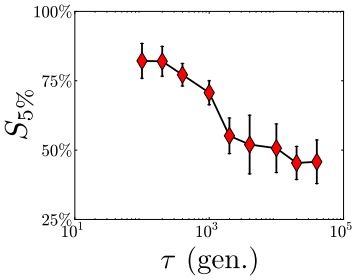

FIGURE 25: Moyenne et écart-type sur 10 simulations de la proportion de couplages ayant un coût supérieur à 5% de l'activité allostérique totale pour une population évoluant dans une période τ fixée.

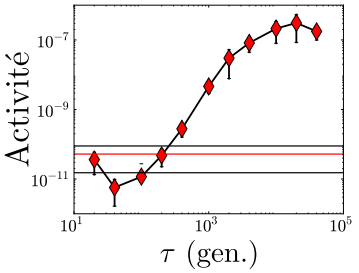

FIGURE 26: Moyenne et écart-type sur 10 simulations de l'activité maximale atteinte par une population évoluant dans une période τ fixée. Les lignes horizontales rouges et noires indiquent les valeurs prises pour le modèle nul.

Dynamique de la population

L'un des avantages de nos simulations est que nous pouvons aussi suivre la dynamique de la réponse d'une population dans cet environnement variable.

Nous nous intéressons ici à la dynamique rapide d'adaptation après un changement d'environnement, c'est-à-dire l'évolution de l'activité moyenne de la population entre deux modifications successives de la fonction recherchée. Une question plus intéressante serait de regarder le temps mis par la population pour s'adapter à un environnement fluctuant, autrement dit : combien de périodes sont elles nécessaires pour que l'évolution produise des protéines parcimonieuses à partir d'une population de protéines dépensières ? Nous pensons cependant que notre modèle n'est pas adapté pour cela car il ne présente pas la frustration qui rend cette dynamique intéressante, les mutations conduisant de manière pour ainsi dire trop directe vers une solution optimale.

La dynamique d'adaptation rapide s'effectue généralement en deux phases : la première rapide, de l'ordre de 50 à 100 générations selon la période τ, consiste en la modification de quelques couplages essentiels pour s'adapter à la nouvelle fonction et dans l'expansion au sein de la population des individus les plus adaptés à ce nouvel environnement. La seconde phase, plus lente, se déroule donc avec une population à nouveau homogène dans laquelle l'évolution améliore lentement les couplages ayant été perturbés lors du changement d'environnement.

Ces deux phases peuvent être plus ou moins faciles à discerner selon l'algorithme génétique et les paramètres utilisés mais nous pensons, de par les mécanismes qui les produisent, qu'elles correspondent à une dynamique proche de la dynamique réelle des protéines. Les expériences d'évolutions dirigées se sont pour l'instant concentrées sur des échelles bien plus importantes (plantes, micro-oragnismes, etc.) [147] et commencent seulement à s'intéresser à la dynamique d'adaptation des protéines [148], nous espérons donc bientôt comparer ces prédictions avec des données expérimentales.

La figure (27) montre cependant que ces deux phases sont généralement difficiles à distinguer au cours d'une seule réalisation car la dynamique, quoique présentant une moyenne relativement nette, est brouillée par une très forte stochasticité comme en témoigne les (très) larges barres d'erreur. Ceci montre aussi la difficulté pour caractériser la dynamique d'adaptation d'une population, on a en effet très vite un bruit important (dès les 15 premières générations sur la figure (27)).

147. Santiago F Elena and Richard E Lenski. Microbial genetics: Evolution experiments with microorganisms: the dynamics and genetic bases of adaptation. *Nat Rev Genet*, 4(6):457–469, June 2003

148. P A Romero and F H Arnold. Exploring protein fitness landscapes by directed evolution. *Nat. Rev. Mol. Cell Biol.*, 10:866–876, 2009

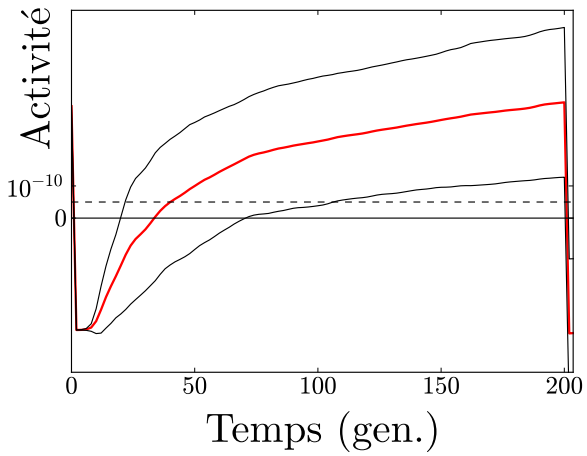

Figure 27: Moyenne et écart-type sur 100 simulations de l'activité moyenne d'une population après un changement d'environnement dans le cas d'un algorithme de type *élite*. La ligne pointillée indique l'activité minimale permettant de confirmer une adaptation de la population. Les chutes brutales de l'activité en début et fin de graphe indique les changements d'environnements.

Un système robuste au détail

> One should absorb the colour of life, but one should never remember its details. Details are always vulgar.
>
> Oscar Wilde, *The Picture of Dorian Gray*

Avant de poursuivre notre analyse, il nous faut prendre un peu de recul pour analyser quels sont les éléments cruciaux de notre modèle et quels en sont les points de détail. Nous voudrions en particulier montrer que le choix précis de la fonction de fécondité ou du plan d'évolution ne modifie pas les résultats fondamentaux de notre simulation. Par ailleurs, le choix exact de l'implémentation de notre modèle n'est jamais essentiel et relève plus d'un caractère pratique que d'une réelle nécessité théorique.

Nous montrerons dans ce chapitre que l'on peut modifier les différents paramètres et même changer complètement la nature du modèle sans perdre le phénomène de concentration des contraintes en environnement fluctuant. Nous terminerons par un résumé des éléments clés qui conduisent à un tel phénomène.

Test

Afin de nous lancer dans les comparaisons, élaborons un test rapide permettant de déterminer si un paramètre est crucial ou non. Les deux résultats essentiels de notre modèle sont :

i un génotype utilisant entièrement ses possibilités lors d'une évolution en environnement constant

ii une évolution dans un environnement variable produit au contraire un génotype parcimonieux.

La valeur précise de la concentration de notre génotype pour une période τ fixée peut éventuellement dépendre des détails de notre modèle. En particulier, un plan d'évolution additif ne semble pas un choix très judicieux et est donc susceptible de présenter certaines caractéristiques pathologiques.

Il est cependant essentiel que la vérification du point ii- ne se fasse pas au dépend excessif de l'activité allostérique. On cherchera donc une valeur de τ présentant une concentration indéniable et une activité α supérieure d'au moins un ordre de grandeur au modèle nul.

À moins qu'il n'en soit spécifié autrement, l'ensemble des paramètres utilisés lors de nos simulations sera toujours le même :

$$\begin{array}{rcll} N & = & 10^3 & \text{taille de population,} \\ \mu & = & 10^{-5} & \text{taux de mutation,} \\ T & = & 5.10^5 \text{ gen.} & \text{temps de simulation,} \end{array} \tag{38}$$

et nous utiliserons un modèle de mutation uniforme ainsi qu'un algorithme de type *élite*.

Fonction de fécondité

Le choix de la fonction de fécondité ne doit pas influencer nos résultats. Dans l'idéal tout algorithme génétique devrait permettre d'exhiber une forme de concentration lors d'une évolution en environnement variable. Nous verrons que même un algorithme de Monte-Carlo présente ce type de résultats ce qui laisse entrevoir la généralité d'une telle propriété.

Notre premier test sera celui de l'algorithme de *sigma-scaling* décrit dans notre section sur les algorithme génétiques (p. 36). Son principal défaut vient d'un recourt abondant à des nombres aléatoires (ce qui ralentit considérablement la simulation) et son manque de contrôle sur la taille de population. Si cette dernière est en effet constante en moyenne, elle reste soumise à des fluctuations qui peuvent éventuellement faire échouer la simulation. Cependant, quand ce n'est pas le cas, cet algorithme est très performant en environnement variable. En effet, un génotype très adapté est susceptible d'avoir de nombreux enfants dans la génération suivante et de prendre le pas sur la population plus vite que dans le cas d'un algorithme de type *élite*. Or dans un environnement variant rapidement, cet effet est crucial pour le développement d'un génotype adapté au sein la population.

Pour une population de taille 10^3, il fallait avec un algorithme de type *élite* au moins 10 générations pour qu'un génotype envahisse cette dernière, ce qui explique les piètres résultats de notre simulation pour $\tau < 200$. Hormis ces détails, cette nouvelle stratégie reproduit fortement les résultats présentés ci-dessus comme le montre al figure (28). C'est d'ailleurs la méthode qui avait été choisie lors de la rédaction de l'article comme nous l'avons fait remarquer précédemment.

Une forme d'implémentation radicalement différente de fonction de fécondité est la stratégie dite de *tournoi* [149]. Un duel est défini comme la compétition de deux individus pris au hasard dans la population. Le plus faible des deux (en termes d'activité) est alors remplacé par une version mutée du vainqueur ; ces deux copies sont ensuite replacées dans la population. Une génération correspondant à un nombre de duels de l'ordre de la taille de population.

149. M. Mitchell. *An Introduction to Genetic Algorithms*. The MIT Press, 1998

Bien que très différent dans sa construction des algorithmes génétiques utilisés précédemment, ce type de modèle produit des résultats très similaire en terme d'activité et de concentration, comme

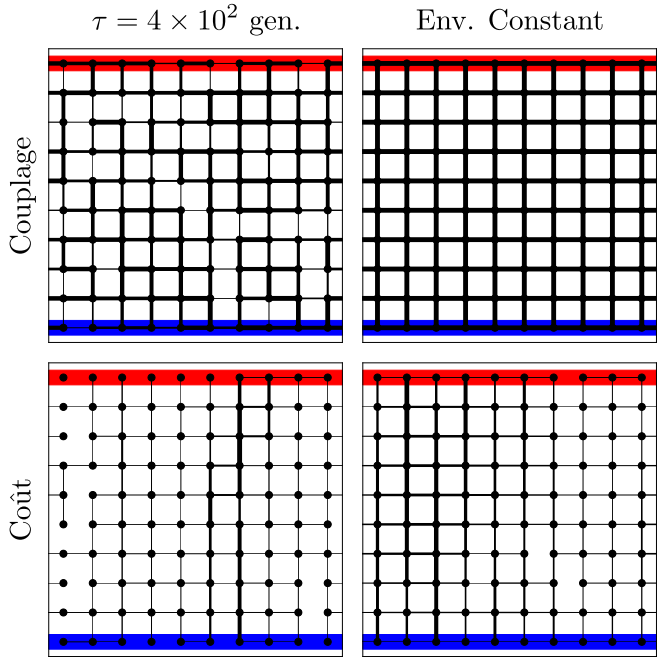

FIGURE 28: Exemple de résultats obtenus en utilisant un algorithme de *sigma-scaling* plutôt qu'une stratégie *élite*. La variante concentrée dispose d'une activité de l'ordre de 5.10^{-10} soit un ordre de grandeur supérieur au modèle nul et est comparable au résultat de la stratégie *élite* dans les mêmes conditions.

le montre la figure (29).

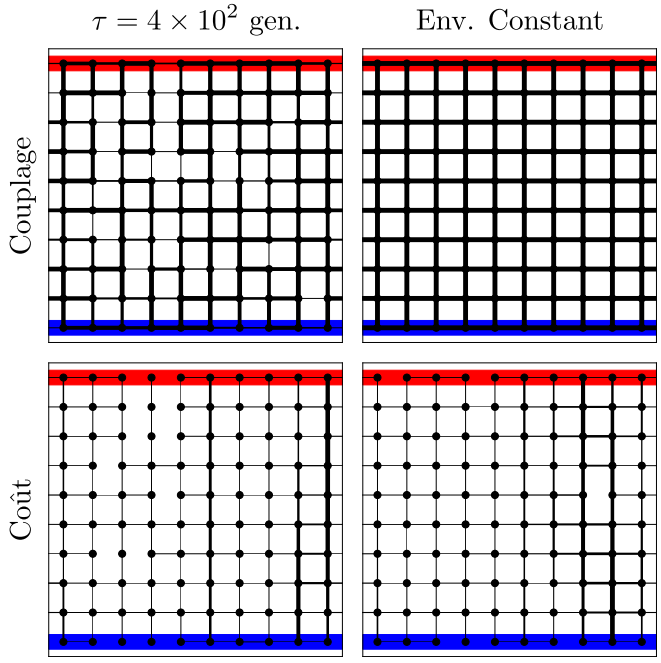

FIGURE 29: Exemple de résultats obtenus en utilisant un algorithme de *tournoi* plutôt qu'une stratégie *élite*. Ici encore, l'activité de la variante concentrée est supérieure d'au moins un ordre de grandeur au modèle nul.

Enfin, la question de savoir si un algorithme génétique est nécéssaire pour observer ce type de résultats ou non vient naturellement à l'esprit, en effet comprendre les différences essentielles entre la dynamique des systèmes biologiques et physiques est cruciale pour ces deux disciplines.

Nous proposons donc l'implémentation d'un algorithme de pseudo Monte-Carlo de la façon suivante [150]. Notre population

150. Cet algorithme est inspiré des *supplementary information* de

Uri Alon, Nadav Kashtan, and Elad Noor. Varying environments can speed up evolution. *PNAS*, 104(34):13711–13716, August 2007

sera constitué d'un seul génotype. Pour déterminer le génotype de la génération suivante, on génère N copies du génotype actuel que l'on soumet à mutation puis on détermine aléatoirement l'une d'entre elles en ajoutant un biais pour sélectionner les copies dont l'activité allostérique est importante. On peut par exemple effectuer un choix aléatoire avec une pondération :

$$p(G(t+1) = G_i) \propto \exp\left(\beta_{\text{MC}} \frac{\alpha_i - \overline{\alpha}}{2\sigma_\alpha}\right), \tag{39}$$

où les notations sont similaires à celles utilisées précédemment ; $\overline{\alpha}$ et σ_α sont respectivement les moyennes et les variances de l'activité au sein de la population et β_{MC} est un paramètre ajustable quantifiant la pression de la sélection naturelle.

Bien qu'il n'y ait plus de notion de population, le phénomène de concentration apparaît aussi en utilisant cet algorithme avec $\beta_{\text{MC}} = 1$ comme représenté dans la figure (30). Notez cependant que cet algorithme n'échantillonne pas toutes les mutations possibles mais seulement une petite partie de l'espace des génotypes autour de $G(t)$, de la même manière qu'un algorithme génétique classique. Il lui faut donc un certain temps pour trouver une mutation efficace mais cette dernière est immédiatement sélectionnée. Il lui est cependant plus difficile de franchir les vallées de basse activité et est donc plus susceptible de se retrouver piégé dans un maximum local. Dans le cas $\beta_{\text{MC}} \to +\infty$, on se retrouve dans un cas similaire à celui de Sélection Forte - Mutation Faible présenté en page 41.

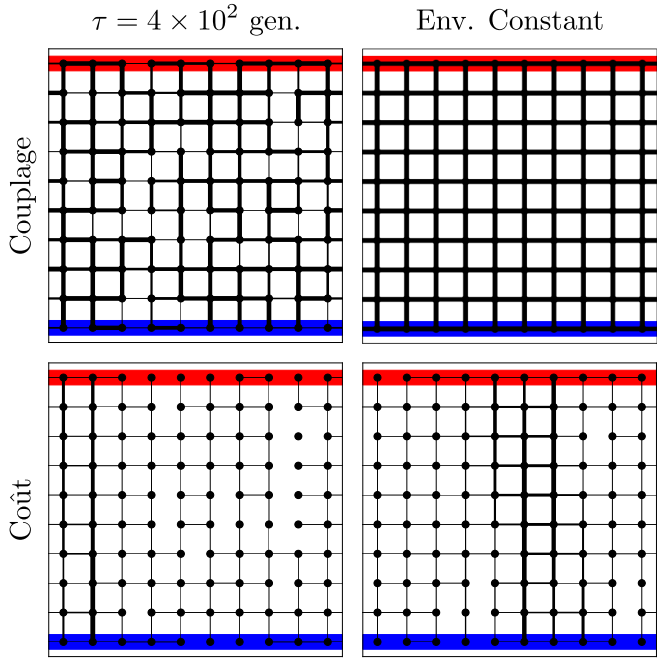

FIGURE 30: Exemple de résultats en utilisant l'algorithme de pseudo Monte-Carlo décrit précédemment avec $\beta_{\text{MC}} = 1$.

Le principe essentiel de la concentration ne réside donc pas dans l'évolution elle-même mais dans la comparaison de différentes échelles de temps. On peut noter trois échelles de temps principales dans notre système :

Le temps minimal d'adaptation τ_a qui est le temps minimal qu'il faut à un système pour s'adapter d'un environnement à un autre. Ici, il est donné par le temps nécéssaire à l'algorithme pour proposer une mutation bénéfique et le temps que cette dernière envahisse la population. C'est le temps en dessous duquel l'évolution n'est pas différentiable du modèle nul,

Le temps de construction τ_c c'est-à-dire le temps qu'il faut au système pour augmenter le nombre de sites impliqués dans la fonction,

Le temps caractéristique de l'environnement τ_e.

On suppose $\tau_a \ll \tau_c$, la question est donc : où se place τ_e ? Si $\tau_e < \tau_a$ le système n'est pas capable de suivre les fluctuations de l'environnement et présente un caractère pratiquement aléatoire ou en tout cas non-adapté. Si à l'inverse $\tau_c < \tau_e$, le système ne réagit pas comme un environnement fluctuant mais comme un environnement constant car sa « mémoire » ne porte pas entre deux changements d'environnement. C'est dans le cas où $\tau_a < \tau_e < \tau_c$ que le système présente une concentration résultant d'une adaptation à un environnement variable.

Si le dernier paramètre est complètement dépendant de l'environnement, les deux premiers sont le résultat à la fois du choix de l'algorithme et de la fonction de fécondité mais aussi du plan d'évolution. Ces dépendances sont potentiellement complexes et mériteraient un approfondissement. Il serait notamment intéressant de comprendre comment le plan d'évolution permet de modifier la géométrie de l'espace des génotypes et de perturber ainsi la dynamique de l'évolution. Nous nous contenterons ici de montrer que les différents choix de plans d'évolution ne perturbent pas notre résultat.

Plan d'évolution

Le plan d'évolution détermine quelles mutations seront proposées à l'évolution afin d'améliorer l'activité allostérique. Il apparaît donc de deux manières dans l'apparition de la concentration. D'une part, c'est lui qui est susceptible de récupérer la fonction après un changement d'environnement et ainsi de diminuer τ_a et τ_c, mais c'est aussi lui qui détermine la facilité à construire *de novo* une nouvelle fonction ou à augmenter la largeur de la région fonctionnelle.

En effet, si le processus de mutation rend difficile d'ajouter des sites aux régions déjà formées, par exemple en ayant tendance à faire dériver vers 0 les couplages sans sélection [151], ce dernier aura tendance à garder constante la largeur de ces régions entre deux changements d'environnement, ce qui facilite une certaine forme de concentration. Un phénomène semblable a déjà été noté par Friedlander *et al.* [152].

Cependant, si pratiquement tous les plans d'évolution possèdent un régime de paramètres dans lequel ils conduisent à l'apparition

151. Comme c'est le cas dans la plupart des processus multiplicatifs puisqu'une fois qu'un paramètre est proche de zéro, les mutations de ce dernier auront par construction des effets faibles.

152. T Friedlander, A E Mayo, T Tlusty, and U Alon. Mutation rules and the evolution of Sparseness and Modularity in Biological Systems. *PLoS ONE*, 2013

de parcimonie, cette région peut être plus ou moins large selon les cas (figure non montrée). Des plans d'évolution pour lesquels τ_a est très important, comme le processus additif décrit au chapitre précédent, exhiberont une très large plage de concentration pour des valeurs élevées de τ_e, signe d'un écartement des pics de haute activité dans l'espace des génotypes munis d'une distance par le plan d'évolution.

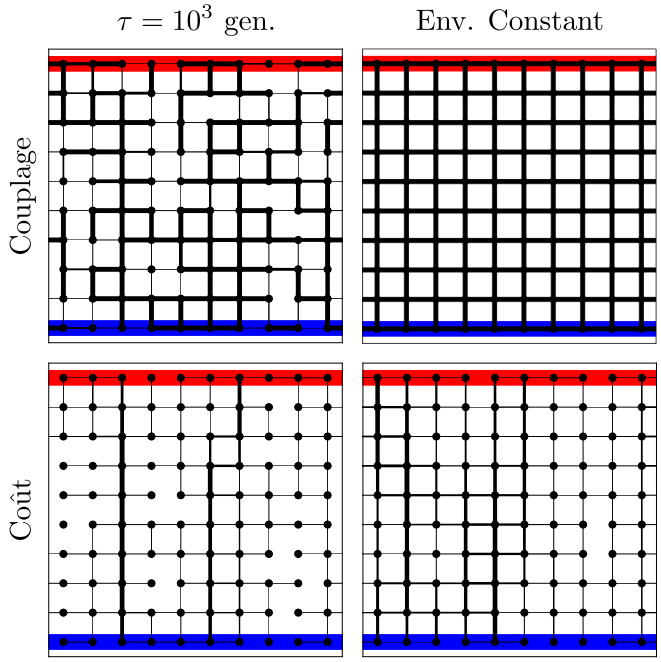

FIGURE 31: Exemple de résultats obtenus en utilisant un plan d'évolution multiplicatif pour lequel les mutations sont représentées par la multiplication du couplage par une variable gaussienne de moyenne unitaire et d'écart-type $\sigma_m = 1.7$.

Les quatre propositions évoquées lors de la description de notre modèle au chapitre précédent montrent ainsi une plage de concentration aux alentours de $\tau_e \simeq 10^3$ générations. Trois de ces plans d'évolution partagent cependant la caractéristique de pouvoir toujours changer le signe d'une interaction en une seule étape – seul le processus additif fait défaut à cette propriété. Ceci est très important car c'est de cette capacité que ces plans d'évolution tirent la possibilité de s'adapter à un nouvel environnement en un faible nombre de mutations. Le phénomène de concentration n'est possible qu'à la condition que l'espace des génotypes présente un chemin simple entre les deux fonctions. Le plan d'évolution détermine la structure géométrique de l'espace des génotypes et détermine donc si un tel chemin existe ou non.

Graphe des mutations

Beaucoup de travaux restent à réaliser afin d'évaluer l'importance du plan d'évolution dans la dynamique de l'évolution. Une piste intéressante pour étudier cet aspect, est d'utiliser comme plan de mutation un graphe de mutations possibles tel que celui de la figure (20). On peut alors analyser la répartition des couplages le long du graphe et comparer le résultat en environnement constant et en environnement variable comme présenté dans la figure (33).

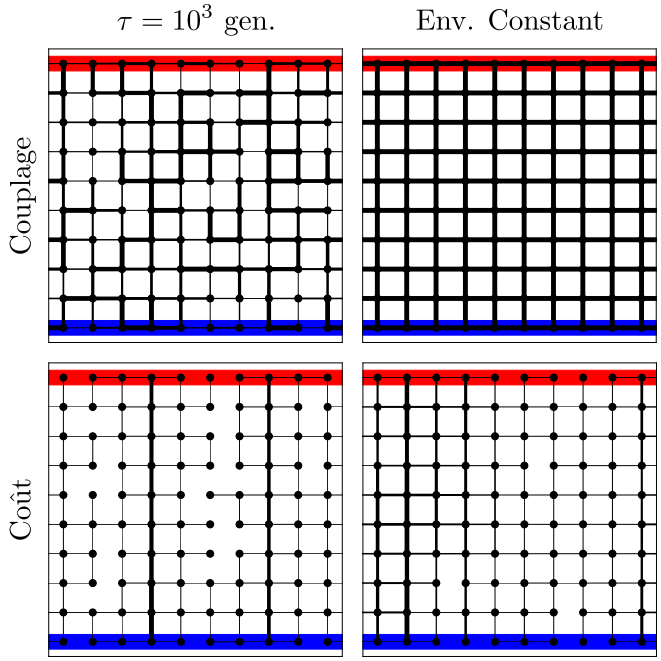

FIGURE 32: Exemple de résultats en utilisant un plan d'évolution additif pour lequel les mutations sont représentées par l'ajout d'une variable gaussienne de moyenne nulle et d'écart-type $\sigma_s = 0.2$.

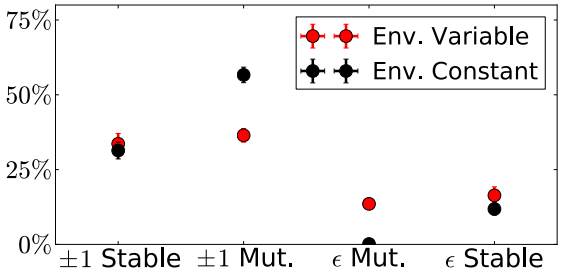

FIGURE 33: Répartition des interactions sur le graphe en fonction de la nature de l'environnement avec $\epsilon = 0.01$. Les barres d'erreur indiquent la déviation standard pour 10 simulations.

De manière étonnante, les couplages fixes ne sont pas spécifiquement plus présents au sein des protéines évoluant en environnement constant mais une différence fortement significative est visible sur la répartition des couplages aisément modifiables qui sont beaucoup plus forts en environnement constant.

Ce résultat laisse ainsi à penser que la capacité d'adaptation d'une protéine ne repose pas fortement sur sur la capacité des acides aminés eux même à s'adapter. Mais l'évolution tire néanmoins profit de cette possibilité puisqu'en environnement constant, pratiquement tout les types mutants sont forts ce qui est loin d'être le cas en environnement variable. Ici encore, l'évolution joue bien plus le rôle d'un bricoleur que celui d'un ingénieur !

Un modèle sur site

Pour terminer ce chapitre, nous proposons de modifier en profondeur le modèle en remplaçant notre génotype continu portant sur les couplages par un génotype discret portant sur les sites. Autrement dit, nous associons à chaque site un entier représentant par exemple un « acide aminé » et déterminerons la valeur des couplages en fonction de la nature des deux extrémités.

Le plus simple pour choisir la valeur des interactions et de supposer que le couplage entre deux sites portant les entiers n_1 et n_2 est donné par une variable aléatoire tirée au début de la simulation. Cette variable doit être comprise dans $[-1, 1]$. Son choix n'est cependant pas crucial, nous avons choisi une gaussienne de moyenne nulle et de variance 0.4 en coupant les queues de façon à rester dans l'intervalle choisi. La matrice est choisie symétrique, mais là encore ce n'est pas important.

Ce modèle est beaucoup plus difficile à équilibrer que les précédents car les couplages sont fortement liés entre eux et changer un couplage ne peut se faire qu'aux dépends de plusieurs autres. Cependant encore une fois, le système montre une concentration après une évolution en environnement variable et une utilisation maximale des couplages dans un environnement constant ainsi qu'il est montré sur la figure (34).

Nous pouvons profiter de ce modèle pour justifier notre utilisation des couplages dans les autres modèles. Nous voudrions pour cela comparer le secteur en terme de sites et celui en terme de couplage. Définir notre secteur par les sites n'est pas difficile puisque nous connaissons exactement la fonction que nous cherchons à implémenter.

Il nous suffit donc, pour déterminer si un site est impliqué dans la fonction, de calculer le coût de toutes les mutations possibles sur ce site et de comparer la valeur de l'activité du génotype originel à la moyenne de toutes les mutations possibles. On peut ainsi définir une quantité normalisée, l'importance du site i pour le génotype G :

$$I(i,G) = \frac{\alpha - \overline{\alpha}}{\sigma_\alpha}, \tag{40}$$

où, comme toujours, nous avons noté $\overline{\alpha}$ et σ_α respectivement la moyenne et la variance de l'activité portant ici sur l'ensemble des mutants possibles au site i. Cette valeur indique donc à quel point l'acide aminé choisit à cette position est effectivement utile pour la fonction recherchée.

On a typiquement trois cas de figure. $I(i,G) > 1$ signifie que le site est particulièrement adapté pour la fonction désirée, $I(i,G) < 1$ signifie que le site va à l'encontre de la fonction recherchée. Cette configuration est particulièrement probable à la suite d'un changement d'environnement et indique un manque d'adaptation. Enfin $|I(i,G)| < 1$ signifie simplement que le site n'est pas soumis à une sélection particulière.

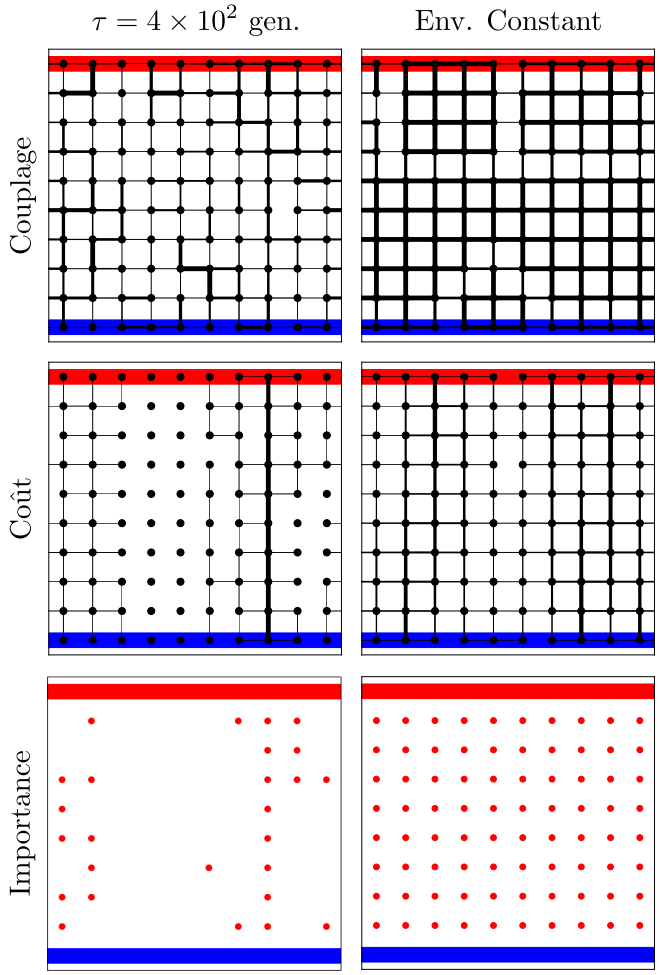

FIGURE 34: Exemple de résultats pour lequel on a utilisé unle modèle d'acides aminés décrit dans le texte. La troisième ligne indique en rouge les sites dont l'importance est supérieure à 1.

Cette méthode permet de recouvrir précisément les sites liés aux couplages essentiels définis par notre fonction de coût. Son principal avantage est de ne pas nécessiter de seuil arbitraire tout en étant plus proche encore des mesures expérimentales réalisées sur des protéines. Comme le montre la figure (34), les deux secteurs recouvrent précisément la même géométrie, ce qui nous indique que notre choix était justifié.

Émergence de modularité

Divide et impera
Philippe II de Macédoine, (359 – 336 av. J.-C)

Dans ce chapitre, nous cherchons à comprendre comment un algorithme génétique peut produire des architectures modulaires alors que dans la plupart des cas, ces derniers mènent à des structures imbriquées [153]. Nous nous demanderons pour cela dans un premier temps pourquoi ces algorithmes évoluent d'ordinaire vers de telles structures avant de montrer que si un environnement variable facilite l'évolution de telles structures, cet effet dépend aussi de la structure des variations utilisées.

153. A. Thompson. *Hadware Evolution: Automatic design of electronic circuits in reconfigurable hardware by artificial evolution.* Springer-Verlag, 1998

Différentes raisons peuvent pousser un algorithme à délaisser la modularité. Premièrement, la plupart des algorithmes utilisent une fonction de coût qui incorpore directement la taille de la description du modèle (par exemple, avec le nombre d'acide-aminés ou le nombre d'opérations requises) afin d'éviter une explosion de la mémoire utilisée par les individus. En effet, cela non seulement ralentirait la simulation mais produirait aussi généralement une grande quantité de d'ADN non codant [154], c'est-à-dire de pièces non utilisées qu'il faudrait ensuite identifier et retirer pour comprendre le fonctionnement réel de l'algorithme [155]. Afin d'éviter cet écueil, il est d'usage de chercher les solutions les plus concises possibles, d'où le coût de taille mentionné ci-dessus. Or généralement, la modularité demandent justement une légère augmentation de la taille de la solution, afin de mieux en séparer les différentes composantes sans pour autant augmenter les performances du système : cette dernière se trouve donc de fait contre-sélectionnée.

154. *junk DNA* en anglais c'est-à-dire l'ADN du grenier : un bric à brac qui ne sert plus mais qu'on garde au cas où !

155. Paul François. Evolving phenotypic networks in silico. *Seminars in Cell and Developmental Biology*, 35:90–97, November 2014

Cependant même en l'absence de coût de taille, par exemple si la taille du problème est fixée comme c'est le cas dans notre modèle, les algorithmes génétiques ne favorisent pas la modularité. La raison en est simplement que les solutions modulaires sont le plus souvent entropiquement défavorisées. S'il existe parmi toutes les solutions un sous-ensemble qui est modulaire, ce dernier ne représente généralement qu'une petite sous-partie de toutes les solutions possibles. C'est-à-dire que le nombre de solutions modulaires N_{mod} est faible devant le nombre de solution non modulaires : $N_{\text{non-mod}} \gg N_{\text{mod}}$. Dans ce cas, la solution de l'algorithme étant

peu ou prou aléatoire parmi toutes les solutions possibles, il est rare – mais pas impossible – qu'il propose une solution modulaire.

D'autres phénomènes, plus spécifiques des problèmes traités, peuvent aussi entrer en jeu. Par exemple dans notre cas, l'une des explications pourrait être qu'il est plus facile d'étendre un canal déjà construit que d'en construire un *de novo* si la fonction ne le nécessite pas. Nous allons désormais montrer un exemple dans lequel notre modèle précédent est modifié pour résoudre une tache modulaire. Si la réponse du système en environnement variable est alors parfois modulaire, nous verrons que le détail de l'environnement participe lui aussi de l'émergence de modularité et peut favoriser cette propriété.

Présentation du modèle

Pour obtenir une forme quelconque de modularité, il est tout d'abord nécéssaire de séparer notre fonction en deux sous-fonctions. Nous garderons ici le cas de l'allostérie mais séparons le site de régulation en deux sites distincts ℓ_2 et ℓ_3. Il y correspond donc deux sous-fonctions dont nous noterons l'activité allostérique α_2 et α_3. L'activité globale que nous utiliserons comme score dans notre algorithme génétique doit tenir compte de ces deux fonctions simultanément :

$$\alpha = \min(\alpha_2, \alpha_3). \tag{41}$$

Dans notre cas, il est aussi nécessaire d'utiliser un environnement fortement variable afin de maintenir la taille du canal petite devant la taille des ligands, nous choisirons donc $\tau = 400$ générations.

Il reste cependant à déterminer comment ces deux sous-fonctions varient temporellement. En effet, les ligands ℓ_2 et ℓ_3 peuvent chacun prendre deux formes mais ces derniers peuvent varier simultanément ou bien de façon décalée en gardant τ constant comme représenté par la figure (35). Ce choix peut paraître anodin, mais il n'en est rien. En effet, c'est ce choix qui aide l'évolution à différencier les deux sous-fonctions comme étant effectivement distinctes puisque le premier environnement ne présente que deux des quatre combinaisons possibles, il n'aide pas l'évolution à séparer les deux tâches.

Ainsi, si deux fonctions se présentent toujours simultanément, même si elles correspondent à des questions différentes, il est plus facile d'implémenter les deux fonctions sur le même canal. Mais si ces fonctions se présentent chacune indépendamment, alors il est plus facile de mettre à jour leur structure réelle… du moins théoriquement. En pratique, ces différences sont généralement une question de probabilité puisqu'il existe toujours la possibilité qu'un algorithme exhibe une solution modulaire par simple chance. La question est donc : « Dans quelle mesure la corrélation entre les sous-fonctions facilite l'évolution de solutions modulaires ? »

La figure (36) présente un exemple de protéine que nous qualifierons de modulaire et un exemple que nous qualifierons de

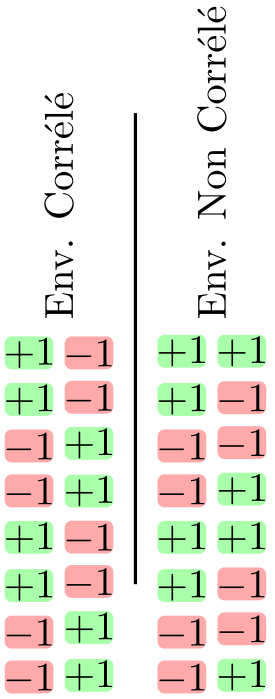

FIGURE 35: Exemples d'environnement, les colonnes de droite et de gauche indiquent dans les deux cas les ligands ℓ_2 et ℓ_3, deux périodes complètes sont représentées ici.

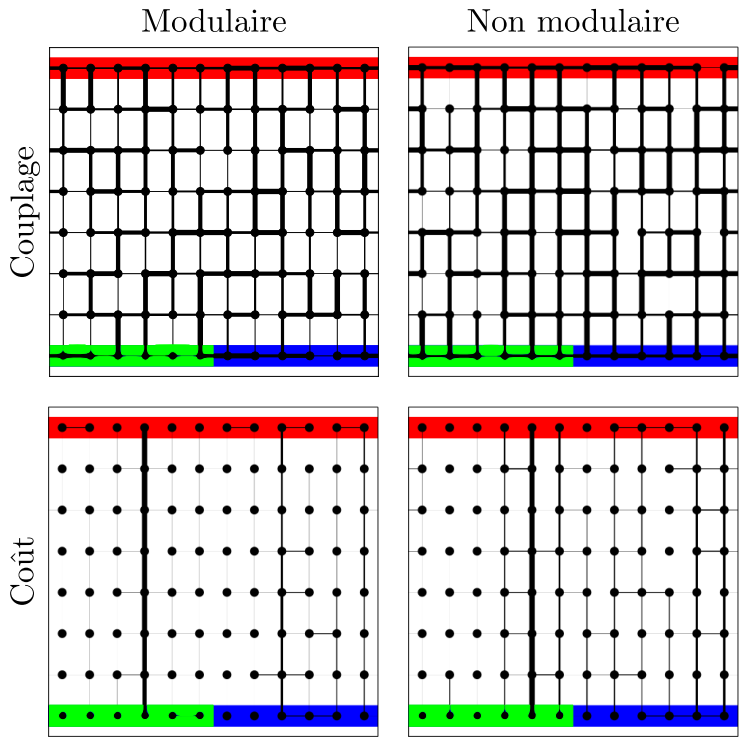

FIGURE 36: Exemples d'architecture de protéine modulaire et non modulaire. Dans les deux cas, l'environnement varie avec une période $\tau = 400$.

non-modulaire. Pour séparer ces cas de manière quantitative, nous utiliserons la procédure suivante. On calcule l'activité restreinte α_2^r (resp. α_3^r) comme l'activité α_2 (resp. α_3) après avoir "éteint" (mis à zéro) tous les couplages correspondant à la partie droite (resp. gauche) de la protéine. On ne calcule donc que la partie impliquée directement avec un ligand. Pour une protéine parfaitement modulaire, les deux fonctions sont indépendantes et on a donc une quantité pratique pour quantifier la modularité :

$$M = 2\frac{\min(\alpha_2^r, \alpha_3^r)}{\min(\alpha_2, \alpha_3)} - 1. \tag{42}$$

Ce ratio va de 0 pour une protéine faiblement modulaire à 1 pour une protéine fortement modulaire.

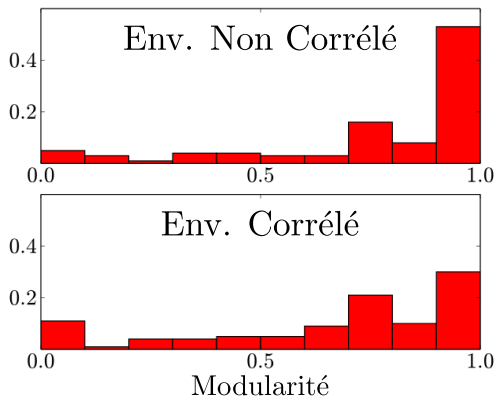

FIGURE 37: Distribution du coefficient de modularité M selon que l'environnement est corrélé ou non pour 100 simulations avec une période $\tau = 400$ dans les deux cas.

Comme le montre la figure (37), le choix de la corrélation entre les changements d'environnement influence l'apparition ou non

de modularité au sein de la protéine. Ceci montre que la manière dont l'on présente les différents environnements induit bien un changement dans la solution qui est alors proposée par l'évolution.

Ces résultats posent de nouvelles questions. À quels éléments de l'environnement cette propriété est-elle reliée ? Dans notre cas, est-ce le fait que l'on présente non plus 2 mais 4 environnements différents qui permet au système de reconnaître la fonction ou bien l'ordre précis dans lequel apparaissent les différents cas as t-il lui aussi une importance pour le développement de l'architecture.

Les liens entre ces questions est celles proches de l'apprentissage automatique seront développés dans les perspectives (p. 107).

Secteurs : Croiser les points de vue

> Il n'y a rien que j'aime moins que de mauvais arguments pour un point de vue auquel je tiens.
>
> Daniel Dennett

Les protéines sont donc des objets particulièrement complexes pour lesquels de nombreux points de vue peuvent être adoptés afin de mettre en valeur telle ou telle propriété. Nous l'avons vu précédemment, des approches structurelles, fonctionnelles ou statistiques apportent des informations et des éclairages complémentaires sur l'architecture et la nature des protéines. Mettre en relation ces différentes approches est donc un objectif important pour la description du vivant. Dans ce chapitre, nous allons donc montrer comment notre modèle simple permet d'illustrer les rapports entre ces angles d'attaque. Pour cela, nous présenterons brièvement ces approches et leurs liens potentiels, puis nous examinerons les informations fournies par ces différentes méthodes sur notre modèle. Enfin, nous montrerons comment notre modèle permet de confirmer certaines hypothèses sur les notions de robustesse et d'évoluabilité.

L'approche structurelle consiste à identifier pour une protéine les éléments qui lui permettent de trouver sa conformation d'équilibre ainsi que d'éventuelles sous structure de cette dernière. D'un point de vue statique, quelles sont les interactions qui favorisent la conformation fondamentale ? D'un point de vue dynamique, dans quel ordre la protéine se replie-t-elle, comment peut-on éviter efficacement les cas de mauvais repliement et existe t-il des groupes d'acides aminés formant des structures rigides au sein de la protéine ?

L'approche fonctionnelle cherche à comprendre le mécanisme qui permet à la protéine de remplir son rôle au sein de l'organisme. Quels acides aminés permettent la catalyse et quels autres favorisent la liaison ? Comment les différentes composantes se modifient-elles lors d'une activité allostérique ? etc.

L'évolution demande d'identifier comment une protéine a pu développer telle ou telle fonction. Elle s'intéresse donc particulièrement aux effets des mutations pour la fonction de la protéine, que ce soit entre les différentes séquences d'une même famille de protéine ou celles d'une protéine donnée au cours d'expériences

d'évolution dirigée. En particulier, on se demandera comment une protéine peut continuer à remplir son rôle malgré les mutations ou comment les mutations permettent à cette dernière de remplir de nouvelles fonctions.

Chacune de ces méthodes fait apparaître un nombre restreint d'acides aminés dotés de propriétés particulières que nous appellerons *secteurs* dans tous les cas. Ainsi, la comparaison statistique des différentes séquences [156] montre l'apparition de secteurs statistiques, c'est-à-dire de groupes de sites présentant des corrélations évolutives fortes au sein d'un groupe mais faibles avec l'extérieur. De même on peut identifier, en quantifiant l'influence de la mutation des différents acides aminés sur la fonction, les secteurs fonctionnels.

156. La méthode SCA que nous avons décrite dans le chapitre dédié (p. 51)

Chacune de ces approches est soumises à des influences différentes (par exemple, la phylogénie influe fortement sur l'analyse statistique des alignements de multiples séquences) si bien que même s'il existe des liens entre les groupes de résidus mis en valeur par ces différentes approches, comprendre la nature de ces liens reste difficile. De nombreux évènements peuvent en effet perturber drastiquement le résultat de l'un de ces points de vue sans pour autant apparaître sur les autres. Il existe toutefois des liens évidents entre ces approches : une séquences qui n'est pas capable de se replier ou de remplir sa fonction sera soumise à une sélection négative, ce qui explique pourquoi les secteurs fonctionnels et structurels sont en partie retrouvés par les méthodes d'étude statistique des séquences. De même qu'il existe des questions importantes : pourquoi les contacts retrouvés à l'aide de la méthode DCA ne se trouvent t-ils jamais entre deux secteurs ? Peut-il y avoir un couplage évolutif sans couplage physique, c'est-à-dire sans contact entre les acides aminés ? Ou plus fondamentalement quel est le lien réel entre les mutations d'une séquence particulière et la statistique de la famille à laquelle appartient cette dernière ?

Pour plusieurs raisons, notre modèle offre un système intéressant pour comprendre les liens entre ces différents points de vue. En effet, il y est possible de définir précisément les différents secteurs et de calculer exactement leur position. Il est également possible de sélectionner une population particulière et de lancer plusieurs simulations différentes partant de ce même ancêtre pour obtenir une évolution artificielle dont nous connaissons exactement les différents évènements.

Au cours de ce chapitre, nous reprendrons le modèle le plus simple dans lequel les mutations sont uniformes et portent directement sur les couplages, c'est-à-dire celui présenté dans le chapitre *Évolution d'un modèle d'allostérie* (p. 73).

Croiser les différents secteurs

Nous cherchons à mettre en correspondance les notions de secteurs fonctionnel, statistique et évolutif. Nous ferons cependant attention à souligner ce qui relève peut-être d'une particularité de notre modèle et ce qui est sans doute une propriété générale des systèmes évolutifs.

Comme nous l'avons souligné précédemment, le secteur fonctionnel peut être défini dans notre modèle en regardant le coût de suppression d'un couplage ; ce qui est effectivement proche de la notion expérimentale de regarder les sites dont les mutations sont les plus significatives en terme de fonction.

Nous aimerions donc désormais faire le lien avec les secteurs tels que définis par l'analyse statistique des séquences génétiques. Pour cela, il faut produire différentes copies de la même population puis les laisser évoluer indépendamment afin de générer un pseudo-arbre phylogénétique très rudimentaire et enfin s'intéresser aux corrélations statistiques entre les génotypes de ces différentes sous-populations. On s'attend alors à trouver des corrélations plus intenses entre les couplages couverts par le secteur qu'entre ceux en dehors de ce dernier.

Pour ce faire, nous avons copié une population que nous désignerons comme la souche parente et avons laisser évoluer 100 populations indépendantes, les populations filles, durant 500 générations, soit un peu plus d'une période. En comparant les valeurs des couplages entre les répliques, nous avons alors pu estimer les corrélations entre les couplages en calculant la matrice d'information mutuelle C_{MI}. Ceci nous permet de définir le secteur principal de corrélation comme le vecteur propre associé à la plus grande valeur propre de C_{MI} (fig. 38 – Corrélations).

Nous pouvons aussi dire un mot rapide au sujet du secteur structurel en utilisant le modèle gaussien élastique dans son objectif originel : regarder le mouvement global des acides aminés d'une protéine. Pour cela, on calcul la matrice de corrélation comme indiqué dans le chapitre approprié (p. 25) et on s'intéresse aux modes dominants de cette dernière, c'est-à-dire au plus grand vecteur propre. On met ainsi en évidence les sites participant au mouvement principal de la protéine, ces derniers se retrouve bien dans le secteur que nous avons découvert à l'aide de nos différentes méthodes (fig. 38 – 1$^{\mathrm{er}}$ Mode).

Toutes ces différentes méthodes mettent en avant le même ensemble de couplage comme montré dans la figure (38) pour les secteurs fonctionnel, statistique et structurel. On remarque cependant que la structure fine de ces différents ensembles varie. Nous pouvons donc, avec une grande confiance, estimer que dans notre modèle, les différents secteurs recouvrent une même réalité physique, mais surtout géométrique. Il existe pour ces individus une région particulière de l'espace remplissant l'essentiel de la fonction : cette dernière présente à la fois de fort couplages, une grande sensibilité

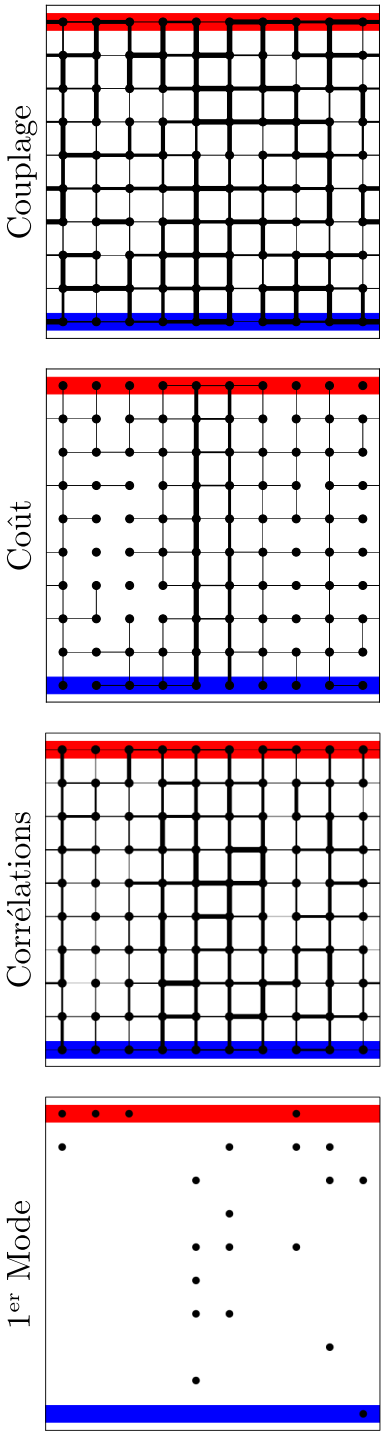

FIGURE 38: Comparaison entre les couplages et les coûts des couplages d'une population mère (secteur fonctionnel), le plus grand vecteur propre de la matrice de corrélation entre les 100 populations filles après 500 générations (secteur statistique) et les 20 sites principaux du plus grand vecteur propre de la matrice de corrélation (secteur structurel).

aux mutations et de fortes corrélations entre ses éléments.

Ces remarques présentent un parallèle direct avec les différents secteurs des protéines naturelles. L'hypothèse la plus forte – au-delà de la structure rudimentaire de notre phylogénie – se situe en effet dans l'utilisation d'interactions qui soit uniquement de courte portée, ici entre plus proches voisins. Or cette hypothèse est couramment utilisé – notamment dans le cadre des modèle gaussiens – et donne d'excellents résultats.

Une différence cependant mérite d'être signalée. En effet, si l'on s'intéresse à la conservation des différents couplages, notre modèle présente un artefact surprenant. C'est dans la zone de plus faible conservation que réside notre secteur tandis que ce dernier constitue au contraire la partie la mieux conservée de la protéine dans le cas naturel. Deux arguments intimement liés peuvent venir expliquer cette divergence : la frustration et la phylogénie.

D'une part nous pensons que ceci peut être le résultat du caractère faiblement frustré de notre modèle. Dans notre modèle en effet, les mutations neutres et les mutations bénéfiques apparaissent avec des échelles de temps similaire : $\frac{1}{\mu N}$. Dans le cas d'un système frustré, les mutations bénéfiques demandent auparavant une première mutation neutre ou défavorable et apparaissent donc avec un taux de mutation beaucoup plus faible de l'ordre (très approximatif) de $\frac{1}{\mu^2 N}$. Pour notre système, cela signifie qu'il est capable d'élargir relativement la taille du secteur fonctionnel, nous sommes obligés, pour observer une forte concentration, de le placer perpétuellement dans un nouvel environnement.

Ceci conduit à un rapport entre environnement et phylogénie particulier dans lequel on a plusieurs périodes de l'environnement (ici deux) depuis la divergence des espèces. Or pour s'adapter, le système sélectionne les mutations bénéfiques, qui se trouvent par définition dans notre secteur. Nous nous attendons donc bien à observer des mutations au sein du secteur en cas de changement d'environnement. Si l'on effectue la même expérience mais qu'on effectue notre analyse statistique avant que les populations filles n'aient connues un changement d'environnement, on s'attend à détecter une forte conservation du secteur ce que confirme la figure (39).

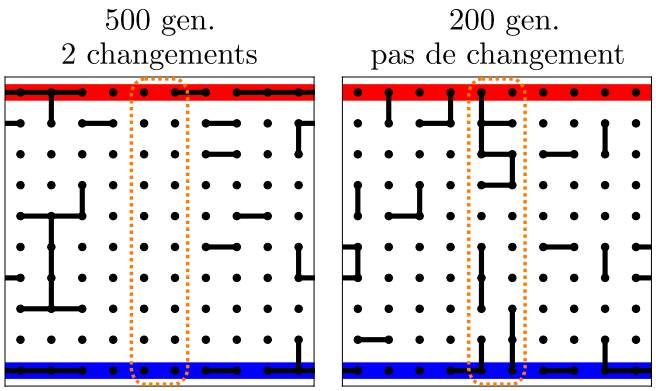

Figure 39: Position des 30 couplages les mieux conservés pour 200 populations descendants d'un même ancêtre. Différence entre une période de 500 générations (soit 2 changements d'environnements) et de 200 générations (aucun changement). Dans les deux cas, le rectangle orange indique la position du cœur du secteur.

Cette remarque nous permet de préciser notre estimateur de l'apparition de secteur $\tau\mu N$. Le terme de mutation μ correspond en pratique aux mutations bénéfiques. Ceci rend difficile une analyse quantitative puisque chacun de ces termes est délicat à estimer.

Hormis donc le secteur évolutif qui présente des particularités par rapport au cas des protéines naturelles, notre modèle permet de se forger une intuition sur les relations entre les multiples définitions des secteurs et confirme qu'il s'agit de différentes facettes d'une même réalité géométrique.

De plus, notre modèle permet d'expliquer pourquoi les différents secteurs d'une même protéine ont tendance à s'exclure mutuellement. En effet, si comme on le pense [157], les différents secteurs assurent différentes fonctions de la protéine telles que le repliement, la liaison, l'allostérie, etc., il y a fort à parier que ces fonctions fluctuent indépendamment dans l'histoire évolutive et induisent donc une séparation des secteurs au niveau de la molécule.

157. Najeeb Halabi, Olivier Rivoire, Stanislas Leibler, and Rama Ranganathan. Protein Sectors: Evolutionary Units of Three-Dimensional Structure. *Cell*, 138(4):774–786, 2009

Évoluabilité et robustesse

Notre modèle permet aussi d'aborder la question ardemment débattue du lien entre robustesse et évoluabilité [158].

Le terme de *robustesse* désigne la capacité d'un système à résister à une modification, souvent une dégradation extérieure ou une défaillance interne. Dans le cas des systèmes vivants, on l'utilise généralement pour désigner la capacité d'un système à conserver sa fonction malgré les mutations génétiques. Un système sera donc dit robuste si la plupart des mutations possibles préservent sa fonction.

Dans le cas de notre système, on notera R la proportion de mutation causant une perte d'activité inférieure à 1%.

L'*évoluabilité* (*evolvability* en anglais) est un terme plus rarement rencontré en dehors de la biologie de l'évolution. Cette notion désigne la capacité d'un système à évoluer et à s'adapter à son environnement. Souvent, ce terme désigne donc à la fois la possibilité de trouver de nouvelles fonctions et celle d'adapter sa fonction à la suite d'un changement d'environnement. Un système sera dit évoluable s'il existe des mutations permettant d'augmenter fortement sa fonction dans un nouvel environnement.

Dans le cas de notre système, on notera E la proportion de mutations provoquant une augmentation de l'activité supérieure à 10% après un changement d'environnement.

Ces deux notions sont centrales pour de nombreux organismes et paraissent souvent contradictoire. En effet, laisser ouverte la possibilité de changer implique souvent de diminuer sa robustesse. Trois types de réponses permettent néanmoins de réunir ces deux notions.

Il est souvent proposé qu'un système robuste soit capable de supporter de nombreuses mutations légèrement défavorables et d'explorer ainsi une plus large portion de l'espace des génotypes, ce qui le rend en pratique plus évoluable [159]. Une autre proposition

158. E van Nimwegen, J P Crutchfield, and M Huynen. Neutral evolution of mutational robustness. *PNAS*, 96:9716–9720, 1999; Andreas Wagner. Robustness, evolvability, and neutrality. *FEBS Letters*, 579(8):1772–1778, March 2005; and P A Gros, H Le Nagard, and O Tenaillon. The Evolution of Epistasis and Its Links With Genetic Robustness, Complexity and Drift in a Phenotypic Model of Adaptation. *Genetics*, 182(1):277–293, April 2009

159. Jeremy A Draghi, Todd L Parsons, Günter P Wagner, and Joshua B Plotkin. Mutational robustness can facilitate adaptation. *Nature*, 463(7279):353–355, January 2010

est de séparer le raisonnement sur le génotype de celui sur le phénotype [160], passant ainsi de l'échelle d'un individu qui ne peut être à la fois robuste et évoluable à celle d'une population qui peut elle combiner les deux propriétés. Ainsi si pour un système donné toute mutation possède un effet important mais que le taux de mutations est suffisamment faible pour que chaque mutant aie eu le temps d'envahir la population si sa mutation est favorable ou de disparaître si elle est délétère, la population peut chercher de manière efficace des solution nouvelles tout en demeurant fermement ancrée sur le meilleur phénotype observé jusqu'à présent. Nous sommes dans le cas de forte sélection mutation faible décrit lors de notre présentation des algorithmes génétiques (p. 41).

Notre modèle suit une troisième hypothèse. En effet, du fait même de la concentration, le système possède de nombreux sites peu importants pour lesquelles les mutations ont un effet négligeable tandis que les mutations au sein du secteur ont un effet important [161], comme montré dans la figure (40). On a ainsi une augmentation du nombre de mutations de faible et de fort effet en séparant géométriquement les deux zones de mutations. De manière amusante, la robustesse apparaît en environnement variable comme un effet secondaire de la concentration qui sert à promouvoir l'évoluabilité. À l'inverse, dans un régime à taux de mutation très fort, le système se concentre pour augmenter sa robustesse et il apparaît une forme d'évoluabilité comme effet secondaire. Les deux propriétés sont alors bien loin d'être antagonistes.

160. Andreas Wagner. Robustness and evolvability: a paradox resolved. *Proceedings of the Royal Society B: Biological Sciences*, 275(1630):91–100, January 2008

161. M M Rorick and G P Wagner. Protein Structural Modularity and Robustness Are Associated with Evolvability. *Genome Biology and Evolution*, 3(0):456–475, January 2011; and A. Thompson. *Hadware Evolution: Automatic design of electronic circuits in reconfigurable hardware by artificial evolution.* Springer-Verlag, 1998

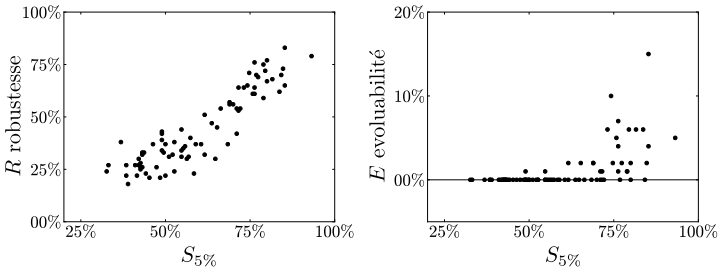

Figure 40: Évolution de la robustesse et de l'évoluabilité en fonction de la parcimonie (ici définie comme dans la figure (25)), pour une large plage de valeur de τ. Ces deux grandeurs sont effectivement corrélées positivement avec la parcimonie de la protéine.

Critiques et développement

> Ne fais pas attention à ce que dit la critique : on n'a jamais élevé une statue à un critique.
>
> Jean Sibelius

Avant de conclure ces travaux sur le développements de parcimonie et de modularité dans des structures évolutives, il nous faut exercer un peu notre esprit critique, ainsi que faire remarquer les nombreux points qu'il reste encore à explorer à l'aide de modèles similaire. Nous présenterons donc dans un premier temps les points faibles de notre système et dans un second temps les améliorations et directions de travail que nous pensons être fructueuses pour de prochaines études.

Critiques

Pour commencer, nous voudrions aborder la question suivante : En quoi ce modèle permet-il d'expliquer la parcimonie des protéines naturelles ? Il n'aura échappé à personne en effet que ce dernier est bien loin de décrire la physique réelle d'une protéine, au sens de ses interactions physico-chimiques avec elle-même et avec le solvant et les ligands. Pourtant, ce modèle partage la structure fondamentale de l'évolution dans laquelle il existe une fonction dépendant de manière complexe de nombreux paramètres qui permet de quantifier le succès, c'est-à-dire ici la capacité d'un génotype donné à posséder des descendants dans les générations suivantes.

Cette fonction, souvent appelée fonction génotype-phénotype (*genotype-phenotype mapping* en anglais) et que nous avons nommée *plan de développement* est l'objet de nombreuses controverse. En particulier, le caractère lisse ou non de cette fonction est l'objet d'une étude approfondie tant au niveau théorique [162] qu'expérimental [163]. Nous entendons ici par « lisse » la notion intuitive de fonction régulière adaptée dans un ensemble discret, autrement dit, une fonction est dite lisse quand la plupart des mutations influent faiblement sur le phénotype et que l'effet d'une mutation dépend faiblement du reste de la séquence – un effet appelé épistasie.

Ces questions cruciales pourraient cependant dépendre fortement de l'échelle à laquelle nous raisonnons, de même que l'apparition de modularité ou plus précisément les phénomènes responsables de l'apparition de cette dernière peuvent être très différents selon les échelles. Un réseau de gène n'est sans doute pas gou-

162. S Kryazhimskiy, G Tkačik, and J B Plotkin. The dynamics of adaptation on correlated fitness landscapes. *PNAS*, 106(44):18638–18643, 2009

163. P A Romero and F H Arnold. Exploring protein fitness landscapes by directed evolution. *Nat. Rev. Mol. Cell Biol.*, 10:866–876, 2009; and Alexandre Dawid, Daniel J Kiviet, Manjunatha Kogenaru, Marjon de Vos, and Sander J Tans. Multiple peaks and reciprocal sign epistasis in an empirically determined genotype-phenotype landscape. *Chaos*, 20(2):026105, 2010

verné par les mêmes types de mutations que la séquence d'une protéine isolée. De même l'aspect lisse ou rugueux du plan de développement pourrait lui aussi dépendre de ces échelles. L'un des problèmes est qu'il est difficile d'obtenir un indicateur fiable de la rugosité d'un paysage, ce dernier étant plus une notion intuitive permettant de se représenter l'évolution qu'un outil permettant son étude quantitative [164].

En l'occurrence cependant, de nombreux indices laissent à penser que notre système est moins frustré – c'est-à-dire que le paysage est plus lisse – que ne le serait une protéine réelle. D'une part à cause des nombreuses sélections négatives s'exerçant sur les protéines [165]. D'autre part, il s'avère que dans notre modèle on peut généralement atteindre le maximum en utilisant des algorithmes de montée de gradient gloutons ce qui laisse à penser qu'il n'y a que peu de maxima locaux et donc un niveau de frustration plutôt faible.

Ceci laisse donc à penser que, dans le cas des protéines, l'extension d'un secteur est difficile à cause des nombreux extrema locaux. On peut donc se demander comment cela modifie nos résultats. Nos simulations montrent qu'en cas de changement d'environnement, la capacité à s'adapter joue finalement un rôle important uniquement sur une courte période. De fait, un génotype adapté prend le pas exponentiellement vite sur le reste de la population. Mais bien que cette phase de compétition soit courte (parfois moins de 5 générations selon l'algorithme utilisé et la taille de la population), elle modifie profondément la dynamique d'adaptation en effaçant les génotypes n'ayant pas réagi suffisamment rapidement. Ainsi, une seule variation d'environnement suffit à épurer la population de tout génotype n'étant pas capable de s'adapter en une génération. Ceci expliquerait pourquoi dans le cas du domaine PDZ, une seule mutation suffit effectivement à modifier l'activité de la protéine d'un spécialiste à un généraliste [166].

Afin de tester cette hypothèse, l'utilisation de modèles physiques provenant des verres de spins connus pour être fortement frustré tels que le modèle NK, ou augmenter la dimension de l'espace (et donc le nombre de contraintes pour un nombre de variables fixe) pourrait apporter un éclairage intéressant au prix d'un éloignement avec la physique du problème des protéines.

164. Sewall Wright. The roles of mutation, inbreeding, crossbreeding and selection in evolution. 1(6):356–366, 1932

165. Dans un organisme vivant, une protéine n'est pas seulement sélectionnée pour répondre à un signal, mais aussi pour ne pas répondre à de nombreux autres !

166. Travaux d'A. Raman, communication personnelle

Un point mérite de plus d'être particulièrement souligné : l'environnement variable n'est qu'une possibilité parmi d'autre permettant d'expliquer la formation de secteurs dans les protéines. Nous l'avons particulièrement mis en valeur par nos choix lors de la construction de notre modèle, mais son influence exacte dans la formation des secteurs pour les protéines naturelles demeure à investiguer.

Par exemple, la géométrie de la protéine peut être particulièrement favorable à l'utilisation d'un groupe d'acides aminés situés

dans une région particulière. On penserait en effet qu'une fonction comme la liaison exclut de fait les résidus situés au plus loin du site actif. Cet effet mérite cependant d'être tempéré. S'il est vrai que les acides aminés importants sont majoritairement situés près du site actif, des études montrent que la mutation d'acides aminées [167] ou la modification de constantes structurelles situés loin du site actif peuvent fortement influer sur l'activité de la protéine [168]. Notez que dans notre cas, le système est placé proche d'une limite de haute température, il est donc fortement sujet à la concentration géométrique, l'information se diluant rapidement le long des liens. Cependant, nous avons choisie une géométrie uniforme et une fonction délocalisée, justement afin de limiter cette influence.

Plus difficile à quantifier, des non-linéarités entre la fonction et le phénotype global peuvent aussi être la source de concentration. Il est ainsi possible que la protéine ne réalise pas la meilleure fonction possible simplement car cette dernière n'est pas nécessaire à l'organisme [169]. L'évolution n'est pas connue pour son amour du travail bien fait et il est probable que certaines protéines soit simplement suffisamment efficaces pour leur tâche sans être parfaitement optimisées. Cette remarque dispose d'au moins un argument solide : la stabilité des protéines. Il est en effet communément admis que de nombreuses protéines – les protéines globulaires notamment – ne sont que marginalement stables. De nombreuses mutations permettent d'ailleurs d'augmenter sensiblement la température de fusion d'une protéine [170]. Il n'est cependant pas très utile à une protéine humaine par exemple d'avoir une température de fusion largement supérieure à 37°C. Et il n'y a aucune raison de penser que l'évolution cherchera à produire des protéines plus stable que nécessaire.

167. Kimberly A Reynolds, Richard N McLaughlin, and Rama Ranganathan. Hot Spots for Allosteric Regulation on Protein Surfaces. *Cell*, 147(7):1564–1575, 2011

168. Thomas L Rodgers, Philip D Townsend, David Burnell, Matthew L Jones, Shane A Richards, Tom C B McLeish, Ehmke Pohl, Mark R Wilson, and Martin J Cann. Modulation of Global Low-Frequency Motions Underlies Allosteric Regulation: Demonstration in CRP/FNR Family Transcription Factors. *PLoS Biol*, 11(9):e1001651, September 2013

169. Ou que le jeu n'en vaille pas la chandelle.

170. Darin M Taverna and Richard A Goldstein. Why are proteins marginally stable? *Proteins: Structure, Function, and Bioinformatics*, 46(1):105–109, December 2001

Améliorations possibles du modèle

Il existe encore de nombreuses directions intéressantes, tant au niveau pratique que théorique, que l'on pourrait explorer en utilisant des modèles similaires.

Frustration. Tout d'abord, le développement d'un modèle présentant davantage de frustration, ou mieux encore permettant de régler la frustration du paysage, serait une suite naturelle à cette étude. Ceci permettrait de mieux comprendre la conservation des secteurs et d'effectuer une analyse temporelle du développement et de la dynamique de ces derniers. Il s'avère cependant qu'un modèle sur sites reproduisant des acides-aminés à l'aide d'une matrice de couplage associée n'est pas suffisant pour obtenir un système frustré. De plus, la frustration est une quantité difficile à quantifier, ce qui complique la recherche dans cette direction. Il est notamment difficile de prouver qu'un système est rugueux et plus encore d'ajuster la rugosité du paysage à un niveau désiré.

Plusieurs pistes méritent néanmoins d'être testées. L'utilisation de plusieurs paramètres, représentant par exemple, la liaison, la

catalyse et le repliement, dans l'évaluation du *plan de développement* pourrait être une piste intéressante mais demanderait un choix questionnable sur la façon de conjuguer ces paramètres. D'une manière plus simple, nous avons dans un premier temps tenté de reproduire une reconnaissance de ligand qui permettrait d'utiliser l'option de la sélection négative mais il s'est avéré que la haute température de notre modèle rendait cette option difficile à implémenter. Un modèle de reconnaissance présentant de nombreuses contraintes pourrait être un premier pas dans cette direction.

Schéma de mutation. L'influence du type de mutation, c'est-à-dire du *plan d'évolution*, sur la structure du paysage et les implications de ces derniers sur l'évolution me paraît être un axe extrêmement intéressant. Toute personne s'étant frotté à la simulation de Monte-Carlo sait combien le temps d'équilibrage d'un système dépend crucialement des mouvements autorisés.

Pour cela, on pourrait dans un premier temps chercher à quantifier les effets de ce paramètre à l'aide de modèles simples comme le graphe de mutations proposés dans la figure (20) pour s'intéresser dans un second temps aux réarrangements plus complets (enjambement, transposons, etc.). Il y a là un défi de taille qui pourrait peut être permettre de comprendre des phénomènes aussi complexes que l'apparition de la reproduction sexuée ou la structure du code génétique.

L'axe principal d'une telle étude devrait reposer sur les liens entre le plan d'évolution et le problème à résoudre. Il n'existe *a priori* aucun plan d'évolution permettant de résoudre facilement n'importe quel problème de même qu'il n'existe pas un algorithme qui soit efficace pour tous les problèmes [171], mais existe t-il un tel plan pour l'ensemble des problèmes biologiques ? Autrement formulé, en quoi le plan d'évolution utilisé à peu de choses près universellement dans le vivant permet-il de tirer des conclusions quand à la nature des problèmes que l'évolution cherche à résoudre ?

171. David H Wolpert and William G Macready. No free lunch theorems for optimization. *Evolutionary Computation, IEEE Transactions on*, 1(1):67–82, 1997

Arbre phylogénétique. Une dernière application de ce type de modèle, notamment de modèles disposant d'une forme de frustration pourrait être l'étude de l'influence de l'arbre phylogénétique sur les analyses de séquences. Il est en effet possible de faire suivre à nos populations de modèles l'arbre de notre choix et d'utiliser les séquences recueillies le long de cette évolution artificielle pour tester, par exemple, les algorithmes de reconstructions d'arbre ou la méthode SCA. La présence d'une contrainte fonctionnelle rapproche en effet ce type de modèles de séquences plus réalistes et fournirait un bon point de référence pour la comparaison de ces algorithmes.

Plus simplement, il est aussi possible de comparer les résultats de l'évolution produits par différentes méthodes de ségrégations et de tester ainsi la notion de séparation géographique et le temps de divergence entre des populations artificiellement séparées.

Troisième partie

Perspectives

Perspectives

> I hear you say "Why ?" Always "Why ?" You see things ; and you say "Why ?" But I dream things that never were ; and I say "Why not ?"
>
> George Bernard Shaw, *Back to Methuselah* 1921

Deux pistes ont été évoquées dans ce tapuscrit qui mériterait une étude approfondie en elle même : les algorithmes d'apprentissage et l'intrication des différentes échelles sur lesquelles agit l'évolution. Il n'est pas question ici d'y répondre mais de montrer les liens entre cette thèse et ces domaines en tentant de souligner pourquoi ces thématiques de recherches nous paraissent intéressantes tant pour la compréhension de l'évolution que pour enrichir les communautés dont sont issues ces questionnements.

Évolution et Algorithme d'apprentissage

Cette section développe les questions soulevées lors de notre chapitre sur la modularité (p. 91). Nous avions fait remarquer que la façon dont nous présentions les différents environnements, ici les associations de ligands, influait sur la reconnaissance par l'évolution des différentes sous fonctions composants la fonction globale.

Intuitivement, ce résultat se comprend bien puisque si deux évènements sont fortement corrélés dans le temps, même s'ils sont issus de deux causes distinctes, il y a un intérêt à les considérer comme liés et à répondre aux deux simultanément. Évidemment ceci n'est vrai que tant que les deux phénomènes continuent de paraître liés. Dès lors que la corrélation disparait, notre réponse sera dès lors inadaptée.

Ce type d'effet peut apparaître au niveau d'une protéine mais aussi d'un réseau de gène si l'organisme suit un cycle de vie particulier comme dans le cas d'*E. Coli* [172] entre autre. Et plus généralement, des effets étrangement similaires apparaissent dans la plupart des systèmes d'apprentissage, on peut penser au conditionnement pavlovien par exemple. En fait, comprendre la structure statistique d'un problème est sans doute la base de tout processus d'apprentissage.

Ce type d'observation pourrait avoir des liens imprévus avec le domaine des algorithmes d'*apprentissage automatique* (*machine learning* an anglais) qui connait une forte explosion ces dix dernières années. L'idée principale de l'apprentissage automatique n'est pas

172. Amir Mitchell, Gal H Romano, Bella Groisman, Avihu Yona, Erez Dekel, Martin Kupiec, Orna Dahan, and Yitzhak Pilpel. Adaptive prediction of environmental changes by microorganisms. *Nature*, 460(7252):220–224, September 2009

nouvelle, il s'agit de chercher à faire apprendre à un système polyvalent disposant de nombreux paramètres les bons réglages pour résoudre une tache particulière. On utilise souvent un système constitué d'un grand nombre d'éléments permettant des calculs simples tel que le neurone présenté en figure (41). Mis bout à bout ces éléments forment un réseau de neurones susceptible de résoudre de nombreuses tâches mais exhibant un grand nombre de paramètre dont l'ajustement est difficile. D'où la nécessité de l'« entraîner » en lui présentant un grand nombre de cas et en ajustant les paramètres au fur et à mesure pour corriger les erreurs.

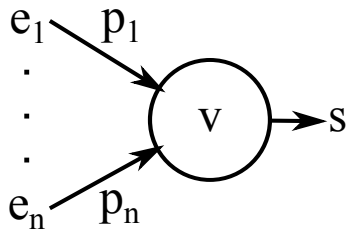

FIGURE 41: Schéma d'un neurone tel qu'utilisé en apprentissage automatique. e_i sont les paramètres d'entrée, s la sortie et p_i et v sont les paramètres internes. On prend généralement une fonction de la forme : $s = \tanh\left(\sum_i e_i p_i - v\right)$ mais une fonction de HEAVISIDE pourrait aussi être choisie.

Pour des raisons de simplicité du calcul, on organise généralement le réseau en couches de telle sorte que les sorties de la couche n forment les entrées de la couche $n+1$. Les capacités d'un tel réseau peuvent dépendre fortement du nombre de couches utilisées [173]. Malheureusement, il est très difficile de remonter aux sources d'une erreur sur plusieurs couches si bien que jusque récemment, les applications se bornaient à des réseaux de 3 ou 4 couches.

En 2006, HINTON *et al.* [174] montrèrent qu'en utilisant une première phase d'apprentissage non supervisé, on obtient un premier jeu de paramètre que l'on pourra par la suite corriger par des méthodes classiques afin d'obtenir un réseau fonctionnel. Cette méthode étant efficace même pour des réseaux dits profonds (*deep* en anglais) c'est-à-dire de 5 couches ou plus. Tout aussi intéressant, le "Curriculum Learning" [175] cherche à reconstruire progressivement la hiérarchie du problème au sein du réseau en choisissant l'ordre dans lequel les éléments de l'ensemble test sont présentés : en partant des cas les plus simples pour aller vers les plus complexes.

Le succès de ces deux approches semble reposer sur l'apprentissage des invariants du problème [176]. En permettant au système de comprendre non pas les réponses à apporter mais la structure même de la question, on augmente ainsi sa capacité à extrapoler des réponses en dehors de l'ensemble d'entrainement.

Notre problème présente de fortes similitudes – toute proportion gardée – avec ces exemples d'apprentissage automatique. Il s'agit de trouver le meilleur jeu de couplage pour calculer la réponse appropriée, la liaison à un ligand cible, à une entrée donnée, la liaison du régulateur. La modularité peut dans ce cas se lire comme la capacité à déterminer la structure du problème, ici sa séparation en deux sous fonctions. Or nous avons montré que cette modularité dépendait de la statistique de l'environnement c'est-à-dire de l'ensemble d'entraînement dans le cas de l'apprentissage automatique.

Dans quelle mesure nos conclusions sur les protéines sont elles valables pour des modèles utilisés dans le cadre de l'apprentissage automatique ? Certain choix de l'ensemble d'entraînement produisent-ils une forme de concentration des contraintes sur une sous partie du réseau de gène ce qui permettrait de mieux comprendre l'architecture de ces derniers ? Existe t-il des géométries particulières – ici au sens de connexions entre les différents élé-

173. Johan Hastad. Almost optimal lower bounds for small depth circuits. In *Proceedings of the eighteenth annual ACM symposium on Theory of computing*, pages 6–20. ACM, 1986

174. Geoffrey E Hinton, Simon Osindero, and Yee-Whye Teh. A fast learning algorithm for deep belief nets. *Neural computation*, 18(7):1527–1554, 2006

175. Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48. ACM, 2009

176. Ian Goodfellow, Honglak Lee, Quoc V Le, Andrew Saxe, and Andrew Y Ng. Measuring invariances in deep networks. In *Advances in neural information processing systems*, pages 646–654, 2009

ments de bases – permettant de conserver la polyvalence du réseau tout en réduisant considérablement le nombre de paramètres à optimiser ?

En retour, ces algorithmes peuvent nous fournir des informations essentiels sur les différences de fond entre l'apprentissage classique et celui mis en œuvre par l'évolution. Le champ des "Curriculum Learning" en particulier permettrait de conduire à une meilleure compréhension des liens entre la statistique temporel de l'apprentissage et la solution qui en émerge.

Évolution à d'autres échelles

Comme nous l'avons fait remarquer à la fin de la partie précédente, l'évolution d'un réseau de gènes est potentiellement très différente de celle d'une protéine [177]. Deux différences fondamentales différencient en effet fondamentalement ces deux systèmes. D'une part, une protéine est fortement contrainte géométriquement là où un réseau de gène l'est *a priori* nettement moins. D'autre part, les enjambements et autres mécanismes de recombinaison jouent sans doute un rôle clé dans l'évolution des réseaux de gènes tandis qu'ils semblent secondaires pour les protéines et sans doute insignifiants dans le cas des domaines de protéines.

La géométrie d'une protéine, tout d'abord, est logiquement contrainte par de simples effets stériques. Le nombre de voisins d'un acide aminé donné est évidemment limité et les conditions aux bords, c'est-à-dire les interactions avec le solvant, forment des contraintes relativement fortes sur la construction d'une protéine. Ceci peut avoir un rôle important sur l'évolution puisque toute modification sera par essence locale, influençant une région de l'espace avec éventuellement quelques interactions le long de quelques axes privilégiés par la protéine, en suivant les secteurs notamment.

À l'inverse, un réseau de régulation n'est pas soumis à cette contrainte et un gène donné peut – au moins théoriquement – interagir avec un nombre illimité de gènes, même si en pratique les gènes interagissant ensemble auront tendance à rester proches le long de la séquence [178]. Comment cette absence de géométrie se traduit-elle en terme de contraintes évolutives ? Et surtout dans quelle mesure cette hiérarchie protéine/réseau de gène est-elle un simple produit des lois de la physique ou bien le résultat d'une évolution cherchant à résoudre des problèmes complexes dans leurs structures [179] ?

Remarquons toutefois que l'évolution même de mécanisme comme celui du transfert de gènes horizontal (HGT), souvent invoqué pour expliquer l'apparition de modularité dans le génome des bactéries, est probablement le résultat d'un environnement propice. En effet, quel intérêt y aurait-il à faire évoluer un mécanisme permettant de déplacer des blocs de gènes répondant à une certaine fonction entre individu si l'environnement n'était pas lui même "découpé" en objectifs modulaires ?

177. Paul François. Evolving phenotypic networks in silico. *Seminars in Cell and Developmental Biology*, 35:90–97, November 2014

178. Ivan Junier and Olivier Rivoire. Synteny in Bacterial Genomes: Inference, Organization and Evolution. *arXiv*, q-bio.GN, July 2013

179. Pour prendre un exemple quelque peu simpliste, penser à la construction d'un langage en terme de lettres/mots/phrases : les règles imposées sur la construction des mots et des phrases sont relativement différentes mais l'architecture rappelle dans les grandes lignes celle du vivant !

Dans tous les cas, comprendre les différences qualitatives et quantitatives dans les phénomènes d'évolution entre ces deux échelles me parait un élément essentiel de la compréhension de la modularité dans le vivant.

Ce cas de figure est bien connu dans le domaine des algorithmes génétiques. Il est ainsi possible d'évoluer une fonction simple à partir de séquences aléatoires. De même, on peut évoluer des fonctions relativement complexes à partir de blocs d'instructions élémentaires. Mais il est extrêmement difficile de chercher à résoudre une fonction complexe *de novo* même si chaque bloc peut être évolué indépendamment et la fonction totale peut être évoluée à partir de ces blocs indépendant. Ce saut conceptuel demande encore une analyse détaillée et il y a fort à parier que la solution réside dans l'utilisation de fonctions d'évaluation plus élaborées permettant au système de comprendre la hiérarchie et la structure du problème.

En montant encore d'une échelle, l'évolution de réseaux trophiques, c'est-à-dire l'évolution des relations entre les différentes espèces d'un environnement, une branche importante de l'écologie, bénéficierait d'un éclairage évolutif plus approfondie. Là encore comprendre les points communs et les différences entre l'évolution d'un tel réseau et celui du génome me paraît être une direction de recherche intéressante.

Quatrième partie

Appendices

Quelques réflexions sur la notion de fitness

> However, it needs to be said that interpretive problems about the fitness concept are not solved by refusing to say what the term means.
>
> Eliott Sober, 1997

Les lecteurs ayant déjà quelques connaissances sur la notion d'évolution et l'usage des algorithmes génétiques n'auront pas manqué de remarquer l'absence complète du terme de *fitness* dans l'ensemble de ce tapuscrit, souvent remplacé par la vocable français de *fécondité*. Il nous a en effet semblé risqué d'utiliser un terme dont la définition n'est pas unanimement partagée ; ce d'autant plus lorsqu'elle est sujette à d'interminable querelles et débats.

Intuitivement, la *fitness* est une quantité permettant de décrire le succès d'un individu donné au grand jeu de la sélection naturelle. Elle doit donc pouvoir être calculé *a priori* et fournir une bonne évaluation de la seule quantité pertinente à l'échelle de l'évolution : la proportion de descendants de cet individu au sein des générations suivantes.

Si l'on se restreint à une demande non quantitative, on peut se restreindre à la notion suivante de fitness : Étant donné un environnement et deux génotypes A et B, on dira que A dispose d'une meilleure fitness que B si l'évolution des deux génotypes dans l'environnement en question mènent majoritairement à la fixation du génotype A.

Définir la *fitness* comme le nombre de descendants est une erreur fréquemment rencontrée. D'une part, cette quantité ne peut être définie qu'à la mort de l'individu et l'on aimerait pouvoir calculer la *fitness* à la naissance de ce dernier. D'autre part, cette définition rend la théorie de l'évolution par sélection naturelle proche d'une tautologie : sont sélectionnés ceux qui ont le mieux réussi. On aimerait pouvoir définir la *fitness* comme, par exemple, l'espérance du nombre d'enfant associé à un génotype donné dans un environnement donné. Cependant, une telle définition non seulement est complexe à calculer et a un pouvoir descriptif limité, mais elle cache aussi deux erreurs subtiles décrites dans *Conceptual Issues in Evolutionary Biology* [180], la première pouvant être facilement réglée, la seconde étant bien plus ardue.

180. E. Sober & al. *Conceptual Issues in Evolutionary Biology*. The MIT Press, 2006

Tout d'abord, prenons l'exemple de deux génotypes donnés, le premier produit toujours 2 enfants, le second a une chance sur deux d'en produire 1 et autant d'en donner 3. En apparence, ces deux individus sont aussi adaptés l'un que l'autre. Pourtant si l'on lance une simulation d'une telle évolution, le premier dominera systématiquement la population à long terme sitôt que la taille de la population sera restreinte. L'effet est facile à comprendre si l'on raisonne sur une population de deux individus que l'on "coupe" à chaque pas de temps pour conserver une taille constante. Si le génotype aléatoire produit 1 descendant, on a un rapport $\frac{2}{3}$ • $\frac{1}{3}$ entre les deux génotypes, tandis que s'il en produit 3, on a un rapport $\frac{2}{5}$ • $\frac{3}{5}$. En moyenne, le rapport est donc $\frac{8}{15}$ • $\frac{7}{15}$ qui est légèrement avantageux pour le génotype constant.

De fait, l'avantage évolutif d'un génotype dont la distribution du nombre de descendant est donnée par e est mieux représenté par

$$s_e = \overline{e} - \frac{\sigma_e^2}{N_{\text{pop}}}, \tag{43}$$

où $\overline{e}$ est la moyenne de cette distribution, σ_e^2 sa variance et N_{pop} la taille globale de la population comme l'a montré Gillespie [181]. Cet effet peut donc être facilement corrigé en utilisant cette nouvelle définition de la fitness. [182]

Le second souci provient des effets survenants après plusieurs générations. En effet, une mutation augmentant le nombre de descendant immédiats mais ne produisant que des descendants stériles par exemple, ne peut pas être qualifiée de bénéfique. En dehors de cet exemple extrême, de nombreux problèmes peuvent se poser. Par exemple, la possibilité de changer le ratio mâle-femelle chez les espèces sexuées peut être bénéfique à court terme – par exemple si la population est particulièrement dissymétrique – mais délétère à plus longue échéance.

On peut aussi imaginer un génotype qui soit particulièrement adapté à un environnement particulier mais peu capable de s'adapter. Ce dernier pourrait donc dominer temporairement la population tant que l'environnement favorable se maintient puis disparaître progressivement par la suite. Doit on tout de même dire que cette mutation est bénéfique pour le génotype ? Eliott Sober propose pour cela d'introduire deux notions de *fitness*, à court et à long terme pour permettre de discuter de ce type de mutation. L'une des principales difficultés étant que cette définition ferait reposer la valeur de la fitness à long terme sur l'histoire future de l'environnement. Ceci ne permet donc de calculer la fitness d'un individu que bien longtemps après sa mort. L'utilité pratique d'une telle définition est laissée à la réflexion du lecteur.

181. John H Gillespie. Natural selection for within-generation variance in offspring number. *Genetics*, 76(3):601–606, 1974

182. Cela veut aussi dire que dans toute simulation où l'on veut que la fitness détermine précisément la propension du génotype à dominer la population, il n'est pas possible de simplement supposer que la fitness détermine la moyenne de la distribution, qu'elle soit poissonnienne ou concentré sur deux valeurs comme c'est généralement le cas.

Dans le cas, plus simple, des algorithmes génétiques, la *fitness* correspond au score de chaque génotype pour le but recherché,

souvent un ensemble complexe des différents sous-buts correspondant à la tâche globale. Notez que le poids de chacun de ces sous-objectifs n'est pas forcément linéaire... ni même connu ! L'évolution d'un robot cherchant à progresser le plus avant possible dans un labyrinthe fait intervenir de très nombreux sous-problèmes comme l'activation concertée de plusieurs moteurs pour déplacer l'appareil, le traitement des capteurs pour repérer et éviter les murs, une construction mentale progressive de la carte du labyrinthe, éventuellement la recherche de sources d'alimentation, etc. Tandis que le score global : « À quelle distance de l'entrée le robot s'est t-il arrêté ? » est d'une simplicité enfantine.

Cette quantité sert donc surtout comme base de comparaison entre les différents individus lors de l'étape de sélection de l'algorithme. Une fois une telle *fitness* calculée, on utilise n'importe quelle fonction permettant de s'assurer qu'un individu ayant une meilleure *fitness* ait en moyenne plus d'enfants afin d'enrichir la population en solutions efficaces pour le problème donné. Notez qu'il y a donc un glissement d'une notion de *fitness* "naturelle" permettant de quantifier l'efficacité du génotype à se reproduire vers une *fitness* "artificielle" qui indique quel individu doit se reproduire. Ce glissement est dangereux pour les biologistes car dans certains articles traitant justement de la question de l'évolution, le terme peut parfois être ambigu comme ici : « *As a result, the probability of each of the folded states was decided when the overall fitness required for reproduction was quite low* » [183] où, bien que l'article traite clairement de l'évolution des protéines, la "*fitness*" utilisée désigne celle d'un algorithme génétique, sans que cela ne soit mentionné dans le texte.

183. P D Williams, D D Pollock, and R A Goldstein. Evolution of functionality in lattice proteins. *Journal of Molecular Graphics and Modelling*, 19:150–156, 2001

De nouvelles difficultés surgissent lorsque l'on s'intéresse à l'évolution biologique : d'une part car l'environnement n'est plus contrôlé et ne cesse de changer, ce qui pose la question de savoir dans quelle mesure la *fitness* dépend de l'environnement et de la capacité de l'individu à y répondre ; d'autre part car la fonction finale qui détermine le succés d'un individu est implicite et n'est pas aussi simple à définir. Elle pourrait être définie par quelque chose comme : *la capacité à assurer une descendance à long terme*. La mention à long terme indique qu'il n'est ni suffisant ni nécéssaire d'augmenter son nombre d'enfants pour augmenter sa *fitness*. Par exemple, protéger ses enfants est une manière simple d'assurer à ses gènes une meilleure place dans les futures populations.

Une notion fréquemment rencontrée est celle de *paysage de fitness* (*fitness landscape* en anglais). L'idée proposée par WRIGHT [184] était de représenter la fitness comme un axe vertical sur l'espace des génotypes et de considérer l'évolution comme une ascension des populations vers les zones de haute fitness. Les images simples comme celle de la figure (42) sont cependant trompeuses, comme l'indique d'ailleurs la légende originelle. Non seulement la géométrie de l'espace des génotypes est très complexe : il est composé

184. Sewall Wright. The roles of mutation, inbreeding, crossbreeding and selection in evolution. 1(6):356–366, 1932

de nombreuses dimensions, ce qui induit des propriétés difficiles à comprendre sur un simple dessin [185]. Mais de plus, la distance naturelle de l'espace des génotypes est induite par le *plan de mutation* et se trouve elle-même très difficile à quantifier. Or c'est elle qui détermine entièrement la dynamique de l'évolution sur ce paysage. Enfin, la question de l'environnement fluctuant a mené certains groupes à envisager la notion de *paysage vague* [186] (*seascape* en anglais) pour souligner l'importance et la difficulté de l'adaptation à un objectif qui n'est pas nécessairement fixe.

Avant de clore ce chapitre, nous voudrions enfin questionner brièvement la pertinence de l'utilisation de la notion de *fitness* pour une protéine isolée. En effet, la notion de *fitness* n'est définie naturellement que pour un individu – ou un génotype. Pourtant il est souvent intéressant d'étudier l'évolution d'un gène particulier au sein d'une population. Plusieurs propositions ont donc été formulées – pour un résumé précis voir Mills [187] mais il convient tout de même d'attirer l'attention sur l'utilisation extrêmement délicate du terme de *fitness* au sein de la biologie : ce terme recouvre en fait de nombreuses quantités différentes liées entre elle par une intuition commune mais floue.

Une fois n'est pas coutume, nous terminerons par une citation résumant en quelques mots les problèmes soulevés ci-dessus :

> Fitness enters population biology as a vague heuristic notion, rich in metaphor but poor in precision.
>
> Richard Levins, 1968

185. Sergey Gavrilets and Janko Gravner. Percolation on the fitness hypercube and the evolution of reproductive isolation. *Journal of Theoretical Biology*, 184(1):51–64, 1997

186. Ville Mustonen and Michael Lässig. From fitness landscapes to seascapes: non-equilibrium dynamics of selection and adaptation. *Trends in Genetics*, 25(3):111–119, March 2009

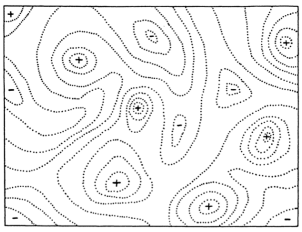

Figure 42: Reproduction du schéma du premier paysage de 'fitness' par S. Wright. La légende en était : *Représentation schématique du champ de combinaison des gènes en deux dimensions plutôt que plusieurs milliers. Les lignes en pointillés indiquent les courbes de niveau de même valeur adaptative.*

187. S. K. Mills and J. H. Beatty. The propensity interpretation of fitness. *Philosophy of Science*, 46:263âĂŞ286, 1979

Lien avec le modèle de verre de spin gaussien

> It is irrelevant that the models may be far removed from physical reality if they can illuminate some of the complexities of the transition phenomena.
>
> Berlin et Kac, 1952.

L'objectif de ce chapitre est de montrer qu'il existe un parallèle simple entre le modèle de réseau élastique développé dans ce tapuscrit (p. 73) et celui de verre de spin gaussien utilisé dans le cadre de notre article [188].

Nous récapitulerons pour cela les grandes lignes de la dérivation du calcul de la fonction de partition d'un modèle de verre de spin dans l'approximation de variables gaussiennes. Nous soulignerons alors les points communs entre les deux approches.

188. Mathieu Hemery and Olivier Rivoire. Evolution of sparsity and modularity in a model of protein allostery. *Phys. Rev. E*, 91(4):042704, April 2015

Partant du hamiltonien ordinaire d'un verre de spin :

$$\mathcal{H}(\sigma_i, J_{ij}) = -\sum_{<i,j>} J_{ij}\sigma_i\sigma_j - \sum_i h_i\sigma_i, \tag{44}$$

dans lequel J_{ij} peut prendre n'importe quel valeur [189]. On cherche une approximation permettant de calculer numériquement la fonction de partition correspondant à un jeu de couplage quelconque de ce modèle dans un temps raisonnable.

189. Contrairement au modèle d'Ising pour lequel J_{ij} est toujours égal à 1.

L'idée est d'utiliser l'approximation gaussienne proposée par Berlin et Kac [190]. Au lieu de considérer que chaque spin σ_i ne possède que deux valeurs ± 1 et de sommer sur toutes les configurations, on va utiliser une distribution gaussienne normalisée pour pouvoir calculer Z analytiquement :

190. T. H. Berlin and Mark Kac. The spherical model of a ferromagnet. *The physical review*, 86:1–15, 1952

$$p(\sigma_i) = \exp\left(-\frac{\sigma_i^2}{2}\right) d\sigma_i. \tag{45}$$

En insérant cette distribution dans l'expression de la fonction de partition, on obtient donc :

$$Z(J_{ij}) = \int \exp(-\beta\mathcal{H}(\sigma_i, J_{ij})) \prod_i \exp\left(-\frac{\sigma_i^2}{2}\right) d\sigma_i, \tag{46}$$

qui peut facilement s'écrire sous une forme matricielle. En notant J la matrice des interactions et σ et h les vecteurs colonnes décrivant

respectivement les spins et les champs extérieurs on a l'expression :

$$Z(J_{ij}) = \int e^{-\beta(-\frac{\sigma^T J\sigma}{2} - h^T\sigma)}.e^{-\frac{\sigma^T\sigma}{2}} d\sigma_i \tag{47}$$

dans laquelle on peut reconnaître la formule de l'intégration matricielle normale :

$$\int e^{-\frac{1}{2}X^T AX + B^T X} dX = \sqrt{\frac{2\pi^n}{\det(A)}} e^{\frac{1}{2}B^T A^{-1} B}. \tag{48}$$

On a donc simplement à remplacer dans la formule précédente :

$$\begin{array}{rcl} A & = & \mathbb{I}_n - \beta J, \\ B & = & h. \end{array} \tag{49}$$

On voit alors combien les fonctions de partition des deux modèles sont proche, la seule différence provenant de la définition de la matrice A [191] :

$$\begin{array}{rcl} A_{\mathrm{GM}} & = & \mathbb{I}_n - \beta J \\ A_{\mathrm{ENM}} & = & \beta(J_r + \sum_j J_{ij})\mathbb{I}_n - \beta J \end{array} \tag{50}$$

191. Pour simplifier nous avons renommer les couplages J, J_r étant toujours le couplage ramenant chaque élément vers sa position

---

En examinant l'équation (49), il est évident qu'il y a un problème semblable à celui de la divergence des réseaux élastiques dans le modèle gaussien de verre de spin. En effet, sitôt que J possède une valeur propre positive – ce qui est plus que probable – il existe une température pour laquelle la fonction de partition n'est plus définie [192].

192. Du moins, cette fonction diverge puis devient complexe, ce qui pour nous revient essentiellement au même.

Pour donner une image physique, le modèle suit une transition de phase qui devrait l'emmener dans une phase de type ferromagnétique, mais comme les spins ne sont pas bornés, il s'en suit une divergence des spins qui se traduit par cette non-analyticité.

Le point qui nous intéresse est cependant la proximité entre le modèle gaussien élastique et le modèle de verre de spin, la seule différence provenant de la diagonale de la matrice à inverser. Il n'existe cependant pas de relation simple entre ces deux modèles même si les limites dans lesquelles $\beta \to 0$ sont identiques pour les deux modèles si l'on impose $J_r\beta = 1$ pour le réseau élastique.

Dans notre cas, on choisit toujours un cas où β est suffisamment faible pour s'assurer une fonction de partition réelle, ce qui indique une température relativement haute et dans tout les cas loin d'être "réaliste" sur le plan physique. On a en effet $\beta J \simeq 0.1$ ce qui en remplaçant les constantes de raideur par l'énergie typique d'une liaison hydrogène de l'ordre du kJ.mol^{-1}, on obtient des température de l'ordre de 10^3 K ce qui est... beaucoup. Ceci explique la tendance de ces systèmes a privilégier les chemins très courts entre les deux ligands et rend difficile leur utilisation pour effectuer des taches plus complexes telles que la reconnaissance de ligands. C'est malheureusement le prix à payer pour obtenir un système demandant un temps de calcul le plus faible possible.

Bibliographie

[1] F Crick and J Watson. Molecular structure of nucleic acids. *Nature*, 1953.

[2] International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature*, 2004.

[3] A K Dunker, P Romero, Z Obradovic, and E C Garner. Intrinsic protein disorder in complete genomes. *Genome Informatics*, 2000.

[4] R J Ellis and S M Hemmingsen. Molecular chaperones : proteins essential for the biogenesis of some macromolecular structures. *Trends in biochemical sciences*, 1989.

[5] H Jane Dyson and Peter E Wright. Intrinsically unstructured proteins and their functions. *Nature Reviews Molecular Cell Biology*, 6(3) :197–208, March 2005.

[6] J B Russell and G M Cook. Energetics of bacterial growth : balance of anabolic and catabolic reactions. *Microbiological reviews*, 1995.

[7] Z Wang, M Gerstein, and M Snyder. RNA-Seq : a revolutionary tool for transcriptomics. *Nat Rev Genet*, 2009.

[8] C B Anfinsen and E Haber. Studies on the reduction and re-formation of protein disulfide bonds. *J Biol Chem*, 1961.

[9] C B Anfinsen. Principles that govern the folding of protein chains. *Science*, 1973.

[10] Stefano Piana, Kresten Lindorff-Larsen, and David E Shaw. How Robust Are Protein Folding Simulations with Respect to Force Field Parameterization ? *Biophysical Journal*, 100(9) :L47–L49, May 2011.

[11] Stefano Piana, Alexander G Donchev, Paul Robustelli, and David E Shaw. Water Dispersion Interactions Strongly Influence Simulated Structural Properties of Disordered Protein States. *J. Phys. Chem. B*, 119(16) :5113–5123, April 2015.

[12] David E Shaw *et al.* Anton, a special-purpose machine for molecular dynamics simulation. *Commun. ACM*, 51(7) :91, July 2008.

[13] Joseph D Bryngelson and Peter G Wolynes. Spin glasses and the statistical mechanics of protein folding. *PNAS*, 84(21) :7524–7528, 1987.

[14] Bernard Derrida. Random-energy model : An exactly solvable model of disordered systems. *Physical Review B*, 24(5) :2613, 1981.

[15] R Mélin, H Li, N S Wingreen, and C Tang. Designability, thermodynamic stability, and dynamics in protein folding : a lattice model study. *The Journal of chemical physics*.

[16] R N McLaughlin, Frank J Poelwijk, Arjun Raman, Walraj S Gosal, and Rama Ranganathan. The spatial architecture of protein function and adaptation. *Nature*, 490(7422) :138–142, July 2012.

[17] Claus O Wilke, Jia Lan Wang, Charles Ofria, Richard E Lenski, and Christoph Adami. Evolution of digital organisms at high mutation rates leads to survival of the flattest. *Nature*, 412(6844) :331–333, 2001.

[18] Sergio G Peisajovich and Dan S Tawfik. Protein engineers turned evolutionists. *Nature methods*, 4(12) :991–994, 2007.

[19] Monique M Tirion. Large amplitude elastic motions in proteins from a single-parameter, atomic analysis. *Phys. Rev. Lett.*, 77(9) :1905, 1996.

[20] Ivet Bahar, Ali Rana Atilgan, and Burak Erman. Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. *Folding and Design*, 2(3) :173–181, 1997.

[21] T Haliloglu, I Bahar, and B Erman. Gaussian dynamics of folded proteins. *Phys. Rev. Lett.*, 79(16) :3090–3093, 1997.

[22] Mathieu Hemery and Olivier Rivoire. Evolution of sparsity and modularity in a model of protein allostery. *Phys. Rev. E*, 91(4) :042704, April 2015.

[23] Swapnil Mahajan and Yves-Henri Sanejouand. On the relationship between low-frequency normal modes and the large-scale conformational changes of proteins. *Archives of Biochemistry and Biophysics*, 567(C) :59–65, February 2015.

[24] Wenjun Zheng and Sebastian Doniach. A Comparative study of motor-protein motions by using a simple elastic-network model. *PNAS*, 100 :13253–13258, 2003.

[25] Francesco Piazza and Yves-Henri Sanejouand. Long-range energy transfer in proteins. *Phys. Biol.*, 6(4) :046014, December 2009.

[26] G N Cohen. Regulation of enzyme activity in microorganisms. *Annual Reviews in Microbiology*, 1965.

[27] Jacques Monod, Jeffries Wyman, and Jean-Pierre Changeux. On the nature of allosteric transitions : a plausible model. *Journal of Molecular Biology*, 12(1-2) :88–118, 1965.

[28] Abel Garcia-Pino, Sreeram Balasubramanian, Lode Wyns, Ehud Gazit, Henri De Greve, Roy D Magnuson, Daniel Charlier, Nico A J van Nuland, and Remy Loris. Allostery

and Intrinsic Disorder Mediate Transcription Regulation by Conditional Cooperativity. *Cell*, 142(1) :101–111, July 2010.

[29] Rhoda J Hawkins and Tom C B McLeish. Coarse-Grained Model Of Entropic Allostery. *Phys. Rev. Lett.*, 93(9) :098104, 2004.

[30] Hesam N Motlagh, James O Wrabl, Jing Li, and Vincent J Hilser. The ensemble nature of allostery. *Nature*, 508(7496) :331–339, 2014.

[31] D U Ferreiro, J A Hegler, and E A Komives. On the role of frustration in the energy landscapes of allosteric proteins. *PNAS*, 2011.

[32] Thomas L Rodgers, David Burnell, Phil D Townsend, Ehmke Pohl, Martin J Cann, Mark R Wilson, and Tom CB McLeish. ΔΔ PT : a comprehensive toolbox for the analysis of protein motion. *BMC Bioinformatics*, 14(1) :183, 2013.

[33] Thomas L Rodgers, Philip D Townsend, David Burnell, Matthew L Jones, Shane A Richards, Tom C B McLeish, Ehmke Pohl, Mark R Wilson, and Martin J Cann. Modulation of Global Low-Frequency Motions Underlies Allosteric Regulation : Demonstration in CRP/FNR Family Transcription Factors. *PLoS Biol*, 11(9) :e1001651, September 2013.

[34] W. B. Provine. *The Origins of Theoretical Population Genetics*. The University of Chicago Press, 1971.

[35] C. Darwin. *On the origin of species by means of natural selection*. J. Murray, 1859.

[36] Rodolphe Barrangou, Christophe Fremaux, Hélène Deveau, Melissa Richards, Patrick Boyaval, Sylvain Moineau, Dennis A Romero, and Philippe Horvath. CRISPR provides acquired resistance against viruses in prokaryotes. *Science*, 315(5819) :1709–1712, 2007.

[37] Ivana Bjedov, Olivier Tenaillon, Benedicte Gerard, Valeria Souza, Erick Denamur, Miroslav Radman, François Taddei, and Ivan Matic. Stress-induced mutagenesis in bacteria. *Science*, 300(5624) :1404–1409, 2003.

[38] O Rivoire and S Leibler. A model for the generation and transmission of variations in evolution. *PNAS*, 111(19) :E1940–E1949, May 2014.

[39] Kalin Vetsigian, Carl Woese, and Nigel Goldenfeld. Collective evolution and the genetic code. *arXiv*, q-bio.PE, May 2006.

[40] François Jacob. Evolution and tinkering. *Science*, 196(4295) :1161–1166, 1977.

[41] Olivier Rivoire and Stanislas Leibler. The Value of Information for Populations in Varying Environments. *J Stat Phys*, 142(6) :1124–1166, March 2011.

[42] S J Gould and R C Lewontin. The spandrels of San Marco and the Panglossian paradigm : a critique of the adaptationist

programme. *Proceedings of the Royal Society B : Biological Sciences*, 205(1161) :581–598, 1979.

[43] Richard A Neher, Colin A Russell, and Boris I Shraiman. Predicting evolution from the shape of genealogical trees. *eLife*, 3, November 2014.

[44] Jesse D Bloom, Claus O Wilke, Frances H Arnold, and Christoph Adami. Stability and the evolvability of function in a model protein. *Biophysical Journal*, 86(5) :2758–2764, 2004.

[45] S.J. Gould. *Wonderful Life : The Burgess Shale and the Nature of History*. New York : W.W. Norton & Co., 1989.

[46] Ivan G Szendro, Jasper Franke, J Arjan GM de Visser, and Joachim Krug. Predictability of evolution depends nonmonotonically on population size. *PNAS*, (2) :571–576, 2013.

[47] Doeke R Hekstra and Stanislas Leibler. Contingency and Statistical Laws in Replicate Microbial Closed Ecosystems. *Cell*, 149(5) :1164–1173, May 2012.

[48] Melanie Mitchell, John H Holland, and Stephanie Forrest. When will a genetic algorithm outperform hill climbing ? In J D Cowan, G Tesauro, and J Alspector, editors, *NIPS*, pages 51–58. Advances in Neural Information Processing Systems, 1994.

[49] B L Miller and D E Goldberg. Genetic algorithms, selection schemes, and the varying effects of noise. *Evolutionary Computation*, 1996.

[50] Q Pham. Competitive evolution : a natural approach to operator selection. *Progress in evolutionary computation*, pages 49–60, 1995.

[51] J.H. Holland. *Adaptation in natural and artificial systems*. The MIT Press, 1992.

[52] M. Mitchell. *An Introduction to Genetic Algorithms*. The MIT Press, 1998.

[53] S. Nolfi and D. Floreano. *Evolutionary Robotics*. The MIT Press, 2000.

[54] Daniel M Weinreich, Suzanne Sindi, and Richard A Watson. Finding the boundary between evolutionary basins of attraction, and implications for Wright's fitness landscape analogy. *J. Stat. Mech.*, 2013(01) :P01001, January 2013.

[55] Bérénice Batut, David P Parsons, Stephan Fischer, Guillaume Beslon, and Carole Knibbe. In silico experimental evolution : a tool to test evolutionary scenarios. *BMC Bioinformatics*, 14(Suppl 15) :S11, October 2013.

[56] E. V. Koonin. *The logic of chance, the nature and origin of biological evolution*. FT Press Science, 2011.

[57] Christof K Biebricher and Manfred Eigen. The error threshold. *Virus Research*, 107(2) :117–127, February 2005.

[58] C Chothia. Proteins. One thousand families for the molecular biologist. *Nature*, 1992.

[59] H Taketomi, Y Ueda, and N Gō. Studies on protein folding, unfolding and fluctuations by computer simulation. *International journal of peptide and protein research*, 1975.

[60] P D Williams, D D Pollock, and R A Goldstein. Evolution of functionality in lattice proteins. *Journal of Molecular Graphics and Modelling*, 19 :150–156, 2001.

[61] K A Dill. Theory for the folding and stability of globular proteins. *Biochemistry*, 1985.

[62] S Miyazawa and R L Jernigan. Estimation of effective interresidue contact energies from protein crystal structures : quasi-chemical approximation. *Macromolecules*, 1985.

[63] Eugene Shakhnovich, G Farztdinov, A M Gutin, and Martin Karplus. Protein folding bottlenecks : A lattice Monte Carlo simulation. *Phys. Rev. Lett.*, 67(12) :1665, 1991.

[64] Hao Li, Robert Helling, Chao Tang, and Ned Wingreen. Emergence of preferred structures in a simple model of protein folding. *Science*, pages 666–669, 1996.

[65] E Bornberg-Bauer and H S Chan. Modeling evolutionary landscapes : mutational stability, topology, and superfunnels in sequence space. In *PNAS*, 1999.

[66] Y Xia and M Levitt. Roles of mutation and recombination in the evolution of protein thermodynamics. In *PNAS*, 2002.

[67] G Trinquier and Y H Sanejouand. New proteinlike properties of cubic lattice models. *Phys. Rev. E*, 1999.

[68] Jeremy L England and Eugene I Shakhnovich. Structural Determinant of Protein Designability. *Phys. Rev. Lett.*, 90(21) :218101, May 2003.

[69] R Helling, H Li, R Mélin, J Miller, and N Wingreen. The designability of protein structures. *Journal of Molecular Graphics and Modelling*, 2001.

[70] Konstantin B Zeldovich, Igor N Berezovsky, and Eugene I Shakhnovich. Physical Origins of Protein Superfamilies. *Journal of Molecular Biology*, 357(4) :1335–1343, April 2006.

[71] E I Shakhnovich and A M Gutin. Engineering of stable and fast-folding sequences of model proteins. *PNAS*, 90(15) :7195–7199, 1993.

[72] Sharad Ramanathan and Eugene Shakhnovich. Statistical mechanics of proteins with "evolutionary selected"sequences. *Phys. Rev. E*, 50(2) :1303, 1994.

[73] S Saito, M Sasai, and T Yomo. Evolution of the folding ability of proteins through functional selection. *PNAS*, 94(21) :11324, 1997.

[74] Walter Fontana and Peter Schuster. Shaping space : the possible and the attainable in RNA genotype-phenotype mapping. *Journal of Theoretical Biology*, 194(4) :491–515, 1998.

[75] S Dalal, S Balasubramanian, and L Regan. Protein alchemy : changing β-sheet into α-helix. *Nature Structural & Molecular Biology*, 1997.

[76] Ayaka Sakata, Koji Hukushima, and Kunihiko Kaneko. Funnel Landscape and Mutational Robustness as a Result of Evolution under Thermal Noise. *Phys. Rev. Lett.*, 102(14) :148101, 2009.

[77] I N Berezovsky and E I Shakhnovich. Physics and evolution of thermophilic adaptation. *PNAS*, 102(36) :12742–12747, 2005.

[78] A Schug and W Wenzel. An Evolutionary Strategy for All-Atom Folding of the 60-Amino-Acid Bacterial Ribosomal Protein L20. *Biophysical Journal*, 90(12) :4273–4280, June 2006.

[79] Benjamin P Blackburne and Jonathan D Hirst. Evolution of functional model proteins. *J. Chem. Phys.*, 115(4) :1935, 2001.

[80] M Heo, L Kang, and E I Shakhnovich. Emergence of species in evolutionary "simulated annealing". *PNAS*, 106(44) :18638–18643, 2009.

[81] Valentina Tozzini. Coarse-grained models for proteins. *Current Opinion in Structural Biology*, 15(2) :144–150, April 2005.

[82] P S Shenkin and B Erman. Information-theoretical entropy as a measure of sequence variability. *Proteins : Structure, Function, and Genetics*, 1991.

[83] Michael Socolich, Steve W Lockless, William P Russ, Heather Lee, Kevin H Gardner, and Rama Ranganathan. Evolutionary information for specifying a protein fold. *Nature*, 437(7058) :512–518, September 2005.

[84] M Weigt, R A White, H Szurmant, J A Hoch, and T Hwa. Identification of direct residue contacts in protein–protein interaction by message passing. *PNAS*, 106(1) :67–72, 2009.

[85] William Bialek and Rama Ranganathan. Rediscovering the power of pairwise interactions. *arXiv*, q-bio.QM, December 2007.

[86] Faruck Morcos, Andrea Pagnani, Bryan Lunt, Arianna Bertolino, Debora S Marks, Chris Sander, Riccardo Zecchina, José N Onuchic, Terence Hwa, and Martin Weigt. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *PNAS*, 2011.

[87] Steve W Lockless and Rama Ranganathan. Evolutionarily conserved pathways of energetic connectivity in protein families. *Science*, 286(5438) :295–299, 1999.

[88] Vasiliki Plerou, Parameswaran Gopikrishnan, Bernd Rosenow, Luís A Nunes Amaral, Thomas Guhr, and H Eugene

Stanley. Random matrix approach to cross correlations in financial data. *Phys. Rev. E*, 65(6) :066126, June 2002.

[89] Najeeb Halabi, Olivier Rivoire, Stanislas Leibler, and Rama Ranganathan. Protein Sectors : Evolutionary Units of Three-Dimensional Structure. *Cell*, 138(4) :774–786, 2009.

[90] O Rivoire. Elements of coevolution in biological sequences. *Phys. Rev. Lett.*, 2013.

[91] Olivier Rivoire, Kimberly A Reynolds, and Rama Ranganathan. The Structure of Evolutionary Constraints in Proteins. *En Préparation*.

[92] Kimberly A Reynolds, Richard N McLaughlin, and Rama Ranganathan. Hot Spots for Allosteric Regulation on Protein Surfaces. *Cell*, 147(7) :1564–1575, 2011.

[93] Nadav Raichman, Ronen Segev, and Eshel Ben-Jacob. Evolvable hardware : genetic search in a physical realm. *Physica A : Statistical Mechanics and its Applications*, 326(1-2) :265–285, August 2003.

[94] R Calabretta, S Nolfi, D Parisi, and G P Wagner. Duplication of modules facilitates the evolution of functional specialization. *Artificial life*, 2000.

[95] Nadav Kashtan and Uri Alon. Spontaneous evolution of modularity and networks motifs. *PNAS*, 102 :13773–13778, September 2005.

[96] Evan A Variano and Hod Lipson. Networks, Dynamics, and Modularity. *Phys. Rev. Lett.*, 92(18) :188701, May 2004.

[97] M Lynch. The Origins of Eukaryotic Gene Structure. *Molecular Biology and Evolution*, 23(2) :450–468, September 2005.

[98] Santiago F Elena and Richard E Lenski. Microbial genetics : Evolution experiments with microorganisms : the dynamics and genetic bases of adaptation. *Nat Rev Genet*, 4(6) :457–469, June 2003.

[99] P A Romero and F H Arnold. Exploring protein fitness landscapes by directed evolution. *Nat. Rev. Mol. Cell Biol.*, 10 :866–876, 2009.

[100] Uri Alon, Nadav Kashtan, and Elad Noor. Varying environments can speed up evolution. *PNAS*, 104(34) :13711–13716, August 2007.

[101] T Friedlander, A E Mayo, T Tlusty, and U Alon. Mutation rules and the evolution of Sparseness and Modularity in Biological Systems. *PLoS ONE*, 2013.

[102] A. Thompson. *Hadware Evolution : Automatic design of electronic circuits in reconfigurable hardware by artificial evolution.* Springer-Verlag, 1998.

[103] Paul François. Evolving phenotypic networks in silico. *Seminars in Cell and Developmental Biology*, 35 :90–97, November 2014.

[104] E van Nimwegen, J P Crutchfield, and M Huynen. Neutral evolution of mutational robustness. *PNAS*, 96 :9716–9720, 1999.

[105] Andreas Wagner. Robustness, evolvability, and neutrality. *FEBS Letters*, 579(8) :1772–1778, March 2005.

[106] P A Gros, H Le Nagard, and O Tenaillon. The Evolution of Epistasis and Its Links With Genetic Robustness, Complexity and Drift in a Phenotypic Model of Adaptation. *Genetics*, 182(1) :277–293, April 2009.

[107] Jeremy A Draghi, Todd L Parsons, Günter P Wagner, and Joshua B Plotkin. Mutational robustness can facilitate adaptation. *Nature*, 463(7279) :353–355, January 2010.

[108] Andreas Wagner. Robustness and evolvability : a paradox resolved. *Proceedings of the Royal Society B : Biological Sciences*, 275(1630) :91–100, January 2008.

[109] M M Rorick and G P Wagner. Protein Structural Modularity and Robustness Are Associated with Evolvability. *Genome Biology and Evolution*, 3(0) :456–475, January 2011.

[110] S Kryazhimskiy, G Tkačik, and J B Plotkin. The dynamics of adaptation on correlated fitness landscapes. *PNAS*, 106(44) :18638–18643, 2009.

[111] Alexandre Dawid, Daniel J Kiviet, Manjunatha Kogenaru, Marjon de Vos, and Sander J Tans. Multiple peaks and reciprocal sign epistasis in an empirically determined genotype-phenotype landscape. *Chaos*, 20(2) :026105, 2010.

[112] Sewall Wright. The roles of mutation, inbreeding, crossbreeding and selection in evolution. 1(6) :356–366, 1932.

[113] Darin M Taverna and Richard A Goldstein. Why are proteins marginally stable ? *Proteins : Structure, Function, and Bioinformatics*, 46(1) :105–109, December 2001.

[114] David H Wolpert and William G Macready. No free lunch theorems for optimization. *Evolutionary Computation, IEEE Transactions on*, 1(1) :67–82, 1997.

[115] Amir Mitchell, Gal H Romano, Bella Groisman, Avihu Yona, Erez Dekel, Martin Kupiec, Orna Dahan, and Yitzhak Pilpel. Adaptive prediction of environmental changes by microorganisms. *Nature*, 460(7252) :220–224, September 2009.

[116] Johan Hastad. Almost optimal lower bounds for small depth circuits. In *Proceedings of the eighteenth annual ACM symposium on Theory of computing*, pages 6–20. ACM, 1986.

[117] Geoffrey E Hinton, Simon Osindero, and Yee-Whye Teh. A fast learning algorithm for deep belief nets. *Neural computation*, 18(7) :1527–1554, 2006.

[118] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48. ACM, 2009.

[119] Ian Goodfellow, Honglak Lee, Quoc V Le, Andrew Saxe, and Andrew Y Ng. Measuring invariances in deep networks. In *Advances in neural information processing systems*, pages 646–654, 2009.

[120] Ivan Junier and Olivier Rivoire. Synteny in Bacterial Genomes : Inference, Organization and Evolution. *arXiv*, q-bio.GN, July 2013.

[121] E. Sober & al. *Conceptual Issues in Evolutionary Biology*. The MIT Press, 2006.

[122] John H Gillespie. Natural selection for within-generation variance in offspring number. *Genetics*, 76(3) :601–606, 1974.

[123] Sergey Gavrilets and Janko Gravner. Percolation on the fitness hypercube and the evolution of reproductive isolation. *Journal of Theoretical Biology*, 184(1) :51–64, 1997.

[124] Ville Mustonen and Michael Lässig. From fitness landscapes to seascapes : non-equilibrium dynamics of selection and adaptation. *Trends in Genetics*, 25(3) :111–119, March 2009.

[125] S. K. Mills and J. H. Beatty. The propensity interpretation of fitness. *Philosophy of Science*, 46 :263–286, 1979.

[126] T. H. Berlin and Mark Kac. The spherical model of a ferromagnet. *The physical review*, 86 :1–15, 1952.

Mathieu HEMERY **21 octobre 2015**

Sujet : Modèles d'évolution de protéines en environnement variable

Résumé : Cette thèse étudie l'influence des fluctuations de l'environnement au cours de l'évolution sur l'architecture fonctionnelle des protéines.

L'apparition de groupes restreints d'acides aminés – les secteurs, possédant des propriétés particulières tant du point de vue structurel et évolutif que fonctionnel ne trouve en effet pas d'explication simple dans le paradigme classique de la physique des protéines. Nous avons donc choisi d'étudier le rôle de l'histoire évolutive dans la construction de cette architecture particulière et des propriétés qui en découlent.

Nous avons pour cela construit un modèle de protéine fonctionnelle inspiré des modèles de réseaux élastiques, que nous avons soumis à une évolution *in silico* en variant au cours du temps, avec différentes fréquences, la fonction recherchée. Nous avons montré que ces fluctuations induisent une concentration semblable à celle observée dans les protéines et avons pu déterminer les paramètres clés contrôlant ce phénomène. Nous avons finalement abordé le lien entre la statistique temporelle de l'environnement et l'apparition de différents secteurs indépendants.

Mots clés : Évolution, Environnement Variable, Protéine, Secteurs, Modèle Élastique, Allostérie, Parcimonie, Modularité

Subject : Models of proteins evolution in fluctuating environment

Résumé : This thesis studies the influence of an evolutionary fluctuating environment on the functional architecture of proteins.

The appearance of restricted groups of amino acids – sectors, with particular functional, evolutional and structural properties has no simple explanation in the classical paradigm of proteins physics. So we choose to study the role of evolutionary history on the construction of this particular architecture and the resulting properties.

We have thus constructed a model of functional protein inspired by the elastic network models, that we have evolved *in silico* while temporarily varying the targeted function with various frequencies. We have shown that these fluctuations induce a form of sparsity close to that observed in proteins and has identified the key parameters of this phenomenon. We finally investigate the link between the temporal statistics of the environment and the appearance of different independent sectors.

Keywords : Evolution, Fluctuating Environment, Protein, Sector, Elastic Network Model, Allostery, Sparsity, Modularity